DOCUMENT RESUME

ED 131 124 TM 005 863

AUTHOR Mills, Roger; Bryan, Miriam M.
TITLE Testing...Grouping: The New Segregation in Southern Schools?
INSTITUTION Southern Regional Council, Atlanta, Ga.
PUB DATE 76
NOTE 82p.
AVAILABLE FROM Southern Regional Council, 52 Fairlee St., N.W., Atlanta, Georgia 30303 ($2.50, Bulk Rate: $2.00, five or more)

EDRS PRICE MF-$0.83 HC-$4.67 Plus Postage.
DESCRIPTORS *Ability Grouping; Achievement Tests; Aptitude Tests; Educational Legislation; Elementary Secondary Education; *Guides; Intelligence Tests; Norms; Reading Readiness Tests; Reading Tests; Southern Schools; *Standardized Tests; *Student Testing; Testing Problems; Test Interpretation; Test Reliability; Test Reviews; Test Selection; Test Validity

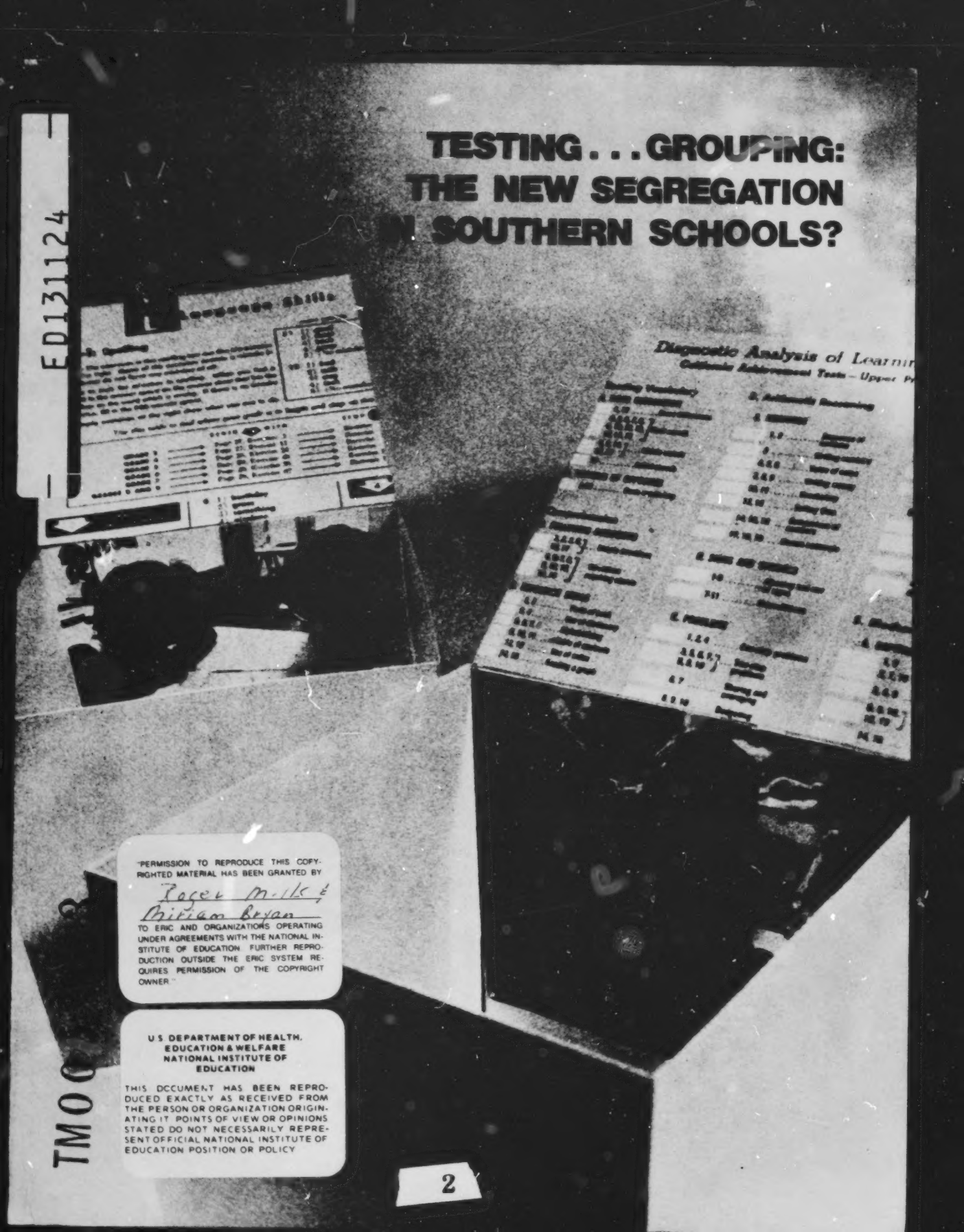
ABSTRACT

When a child takes a standardized test measuring ability or achievement, the results of that test may well determine the kind of education he is going to get. When school personnel select certain tests, administer them, and assign students to groups according to their scores on these tests, are they doing it in such a way that the end results are educationally sound for all children? The purpose of this handbook is to help answer this question. It explains what testing is all about and offers suggestions on how the testing and grouping of children in a school system may be examined. Information is provided in these areas: (1) ability grouping and tracking, (2) testing, (3) how to find out about ability grouping and tracking in a school system, (4) how to begin the move for reform in a system, and (5) ability grouping and the law. Appendices contain a section on the prevalence of ability grouping across the South, "Ability Grouping: Status, Impact, and Alternatives" by Miriam Bryan, descriptions of tests commonly used in ability grouping, and a glossary of measurement terms used in this handbook. (RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED131124

TESTING . . . GROUPING: THE NEW SEGREGATION IN SOUTHERN SCHOOLS?



"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY

Roger Mills & Mitian Bryan

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TESTING . . . GROUPING:
THE NEW SEGREGATION IN SOUTHERN SCHOOLS?

CONTENTS

INTRODUCTION 1

ABILITY GROUPING & TRACKING 2

TESTING 8

HOW TO FIND OUT ABOUT ABILITY GROUPING & TRACKING IN YOUR SCHOOL SYSTEM 20

HOW TO BEGIN THE MOVE FOR REFORM IN YOUR SYSTEM 30

ABILITY GROUPING & THE LAW 39

APPENDICES:

A. *Prevalence of Ability Grouping Across the South* . . . 45

B. *Ability Grouping: Status, Impact, and Alternatives* by M. Bryan. Eric, June, 1971 . . . 49

C. *Descriptions of Tests Commonly Used in Ability Grouping* . . 53

D. *A Glossary of Measurement Terms Used in This Handbook* . . 72

© Copyright 1976 Southern Regional Council

52 Fairlie Street, N.W.
Atlanta, Georgia 30303

Price: $2.50; Bulk rate: $2.00 (5 or more)

ABOUT THE AUTHORS

ROGER MILLS, a staff member and an attorney at the Southern Regional Council, has been active in school desegregation litigation for a number of years. In 1974 he authored *Justice Delayed and Denied: HEW and Northern School Desegregation.*

MIRIAM M. BRYAN, test specialist, was associated for many years with Educational Testing Service, most recently as associate director of the Atlanta Office of that organization. In 1971 she coauthored with Warren G. Findley the volume entitled *Ability Grouping: 1970 Status, Impact, and Alternatives*; and in 1975 she and Dr. Findley coauthored the Phi Delta Kappa fastback entitled *The Pros and Cons of Ability Grouping.*

INTRODUCTION

When your child picks up a soft-leaded pencil at school and fills in certain blanks on an answer sheet to a standardized test measuring ability or achievement, the results of that test may well determine the kind of education he is going to get. The results may determine whether he will be placed in a slow group or an average group or whether he will be recommended for vocational studies or college preparatory studies. School officials frequently use test results to assign students, teachers, and textbooks in such a way that some children go fast and some go slow.

There is nothing wrong with this theoretically since students do proceed at different rates of progress and are able to achieve different levels of success in their studies. The point is, though, that if school personnel select certain tests, administer them, and assign students to groups according to their scores on these tests, are they doing it in such a way that the end results are educationally sound for all children?

The purpose of this handbook is to help you answer these questions. The handbook tells you what testing is all about and offers suggestions on how you may examine the testing and grouping of children in your school system.

ABILITY GROUPING AND TRACKING

A. ABILITY GROUPING

Ability grouping is the practice of arranging groups of students in different sections or classrooms within a grade and assigning different teachers to these groups. The students may be grouped on the basis of (1) how well they do on tests measuring aptitude or achievement, (2) their past grades, (3) the judgment of the school counselor or teacher, or (4) a combination of these things. If achievement tests are principally used to determine the groups to which students are assigned, the practice is sometimes called achievement grouping rather than ability grouping.

If a teacher informally groups students within a regular classroom for different subjects for parts of the day and for varying lengths of time, this is not considered to be ability grouping in the strictest sense. This is frequently the only way in which students who are deficient in a particular skill can get the extra help they need to overcome the deficiency. It becomes ability grouping only if the grouping is not temporary and if the different groups are assigned to different classrooms and to different teachers and are taught at a different pace.

Assigning students according to ability is different from assigning students to special education classes. Before a student can be placed in a special education class, most State Department of Education procedures require that the student be given an individual examination by a physician and by a psychologist and that consent be obtained from the parent. When students are ability-grouped, however, the tests given are more than likely to be group-administered and no consent of the parents is sought. Special education students include those who are physically handicapped with speech, hearing, visual, or other impairments; and those who are mentally handicapped, or, to use the terms employed by educators, are educable mentally retarded (EMR) or trainable mentally retarded

(TMR), or have learning disabilities (LD) or behavior disorders (BD).[1] While these special education students may be separated from the ordinary classroom, they are not considered to be ability-grouped and so are not covered by this handbook.

One example of ability grouping in practice is the testing of students for reading achievement level and permanently assigning high and low scorers to different classrooms in which all subjects are tailored to reading level. Another is the testing of students in several basic subjects and more or less permanently placing them at different levels for each subject. Assignments may be made, for example, so that third graders who have problems with arithmetic are taught with advanced second graders. Unfortunately, once so assigned, these students have little opportunity to proceed at their own pace when they have overcome their problems and, as a result, they never do catch up with other students at their own grade level but tend to remain with the group of their initial assignment.

B. TRACKING

Tracking is the practice of assigning students at the high school level, and often at the middle school level, to certain self-contained curricula such as college preparatory, business, or vocational studies. The courses are frequently themselves divided into honors, general, and remedial sections. Tracking at the secondary level differs from ability grouping at the elementary level because the older students themselves choose their programs of study with the help of their parents and teachers or counselors. Because tracking may imply so much inflexibility that a student may not switch from vocational to college preparatory studies or from a

[1] Educable mentally retarded students usually have IQs in the 50-75 range. Trainable mentally retarded students have IQs in the 35-50 range. Students with learning disabilities are those who have normal IQs and have normal vision and hearing but cannot learn well because they have problems in perception that prevent them from processing information properly. Students with behavior disorders are emotionally disturbed or socially maladjusted.

remedial English course to a general English course, many school officials prefer not to use the word "tracking." It is true, however, that considerations such as class scheduling, many required courses with few electives, and course unit prerequisites severely restrict students from moving in and out of various programs of study or courses so that tracking does in fact occur. Then, too, while it may appear that students freely select their courses of study, closer examination may show that student choice is, in fact, heavily influenced by counselor advice, teacher recommendations, and the desire to be with friends -- all matters determined largely by how the student was ability-grouped in previous grades.

One example of tracking in practice is a high school providing academic, general, and business education courses of study with required subjects such as English, mathematics, social studies, and science divided into basic, regular, and honors offerings. In such a school, business students would take business English and earth science, while academic course students would take English composition and biology and chemistry.

C. EDUCATIONAL ADVANTAGES AND DISADVANTAGES

School personnel who favor ability grouping or tracking argue that it has several advantages. Some of the most frequently made claims include these:

1. It permits students to move at their own learning rates.
2. It allows students to compete on a more equitable basis.
3. It makes curriculum planning easier for the teacher since there is not such a wide range of students to teach.
4. It makes it possible for each student to taste some success.
5. It improves student self-image in the groups that are ahead.

6. It is favored by parents, especially those of more talented students.
7. It makes it possible to have both remedial and enrichment programs.
8. It simplifies scheduling procedures for the administrator.
9. It decreases discipline problems.
10. It permits the more efficient purchase and use of materials.

On the other hand, many educators have found ability grouping and tracking to be actually harmful to students. The most frequently stated disadvantages they cite are these:

1. It reduces, for students placed in groups that are average or behind, the stimulation and leadership provided by the students who are in groups that are ahead.
2. It stifles the socialization process, making the students who are ahead more snobbish and the students who are behind feel like second-class persons.
3. It encourages some teachers to expect less from the students in groups that are behind and then in turn to teach them less.
4. It may result in segregation of students by race or by family income.
5. It fosters unhealthy self-concepts among students placed in groups that are behind.
6. It creates morale problems for teachers assigned to teach the groups that are behind.
7. It destroys the challenge of competition.
8. It may result in the assignment of the least able and least experienced teachers to the groups that are behind.
9. It prevents students who are ahead from becoming sensitive to the problems of other students.

10. It allows students little opportunity for movement throughout the school years as a result of the initial grouping.

11. It results in decreased motivation, especially at the average and below-average levels.

12. It is sometimes based on invalid criteria.

13. It occasionally creates parent pressure to assign a student to classes too advanced for the child.

14. It may concentrate students who have discipline problems.

15. It has not been shown to improve learning and may impede student progress in the higher grades.

16. It implies that a student with low achievement at one grade is a slow learner.

D. RESEARCH FINDINGS

Voluminous literature covering more than sixty years of research concerning ability grouping was reviewed by Warren G. Findley and Miriam M. Bryan in *Ability Grouping: 1970 Status, Impact, and Alternatives.*[1] The authors reached several conclusions.

They reported that ability grouping may adversely affect

---

[1] Historically, the practice of ability grouping arose and flourished from about 1915 until 1935, when progressives motivated by humanitarian concerns decided that slow children ought to be separated from others and allowed to proceed at their own rate to give them a taste of success. In the twenty years following, it fell into disrepute since it could not be justified in terms of clear advantages to students. There was a resurgence of interest in the middle 1950s spurred by the Russian launching of Sputnik and the consequent emphasis on special training for talented students in the areas of science and mathematics. The interest has continued to the present with the proliferation of separate funding programs at both federal and state levels for gifted children and children with mental or physical handicaps, and with the implementation of school desegregation across the South, where students formerly provided with different educational opportunities were combined into single unitary school systems.

student achievement:

> *Briefly, we find that ability grouping. . . shows no consistent positive value for helping students generally, or particular groups of students, to learn better. Taking all studies into account, the balance of findings is chiefly of no strong effect either favorable or unfavorable. Among the studies showing significant effects the slight preponderance of evidence showing the practice favorable for the learning of high ability students is more than offset by evidence of unfavorable effects on the learning of average and low ability groups, particularly the latter. Finally, those instances of special benefit under ability grouping have generally involved substantial modification of materials and methods, which may well be the influential factors wholly apart from grouping.* Findley and Bryan, p. 54.

They reported that ability grouping may adversely affect self-concept, attitudes, and personality traits:

> *The findings regarding impact of ability grouping on the affective development of children are essentially unfavorable. Whatever the practice does to build (inflate?) the egos of children in the high groups is overbalanced by evidence of unfavorable effects of stigmatizing average and low groups as inferior and incapable of learning.* Findley and Bryan, p. 54.

They reported that ability grouping may increase racial and class segregation:

> *In the absence of evidence of positive effects on learning and personal development of children, and in light of negative effects on the scholastic achievement and self-concepts of low ability groups, the tendency of ability grouping to separate children along ethnic and socioeconomic lines must be deemed to discriminate against children from low socioeconomic classes and minority groups. The mechanism may be said to operate primarily in denying the low groups the scholastic stimulation of their more able peers, and by stigmatizing the low groups as inferior and incapable of learning in their own eyes and those of their teachers.* Findley and Bryan, p. 54.

TESTING

A. ABILITY v. ACHIEVEMENT TESTS

Regardless of how students are grouped or tracked, tests of some sort are invariably used in the process. Sometimes the tests are teacher-made classroom tests; sometimes they are developed by teacher committees according to local specifications; sometimes the tests are provided by textbook companies to accompany their own reading or mathematics programs; but usually they are commercially published ability or achievement tests that have been standardized on large student populations through pretesting and norming.

Although there are over 2,000 standardized tests currently being published in this country, most of the tests with which schools are concerned are of two types: ability tests or achievement tests.[1] Ability tests, also known as aptitude or intelligence tests, are designed to measure academic potential. Achievement tests, on the other hand, are designed to measure how much a student has learned in a specific subject and are, therefore, supposed to reflect what has actually been taught. Achievement tests may be for individual courses such as reading, algebra, or biology, or they may be "batteries" or groups of tests that cover a number of subjects. Tests for specific subjects vary somewhat from one to another; but batteries usually include tests on reading, language skills, and mathematics, and may include tests on science and social studies as well. Most test batteries are prepared for different age levels. For example, one level may be for grades 1-3, another for grades 4-6, and so on.

Until recently most standardized achievement tests have been norm-referenced. Recently, there has been considerable enthusiasm on the part of school people for criterion-referenced achievement

---

[1] Brief descriptions of ability and achievement tests most commonly used in student grouping are included in Appendix C.

tests. Criterion-referenced tests differ minimally from norm-referenced tests except that they may focus more strongly on objectives of the local instructional program and evaluation of test results is based on how well the individual student has achieved the locally determined goals rather than on how the achievement of the individual student, the class, the grade, or the system compares with the achievement of individual students and groups beyond the local district. While to date most criterion-referenced tests have been locally constructed and administered and are, therefore, not standardized, several test publishers are currently working on the development of marketable tests of the criterion-referenced type.

Ability tests are designed to measure mental abilities and skills developed over a long period of time. Achievement tests, whether norm-referenced or criterion-referenced, measure skills learned over a shorter period of time. Test critics often maintain that there is little difference between what ability and achievement tests actually measure, and this is more likely to be true than not. Both types of tests have similar content at various age levels and children normally do about as well on one type of test as on the other. Because of these similarities, both types of tests tend to have much the same attributes and the same shortcomings.

B. VALIDITY, RELIABILITY, AND NORMING

Because test results play such an important part in a student's educational development, it is extremely important that the tests do the job they are designed to do. To do their job, standardized tests must be satisfactorily valid and reliable, and should be adequately normed.

The validity of a test is determined by how closely the test matches up with the real world things it is supposed to measure. There are several kinds of validity, three of which are of special significance in school testing situations. Content validity is determined by how well questions on the test relate to the courses of

study being taught. Concurrent validity is determined by how well the test scores match with a student's school grades or, since school grades are not always reliable, with some other indicator to which the test scores may be related. Predictive validity is determined by how well the test scores can predict the future performance of a student.

The reliability of a test is determined by how consistently the test measures student performance, that is, whether a student gets about the same score on a test the second time around. Reliability can be estimated by determining the consistency of test scores by splitting the test in half and comparing performance on both halves; by comparing scores on different, but comparable, forms of the test; and by calculating the degree of stability of scores in the same test or similar tests over a period of time.

When is a test sufficiently valid or reliable? Theoretically, estimates can range from -1.00 to +1.00. Standardized tests never reach these extremes. If the validity coefficient ranges between .40 and .70, it is satisfactory; and if the reliability coefficient ranges between .70 and .90, it is acceptable. These figures are usually stated in the publisher's instruction manual accompanying the tests.

Norms provide a specific comparison group, in relation to which test performance of students may be meaningfully described. A norms group may, for example, represent all fourth graders in the nation or all high school students in a state who have completed a course in American history. Norms for a test are based on a distribution of all the scores on the test made by those in the norms group. The distribution covers the entire range of scores from the lowest to the highest. A "norm," as distinguished from "norms," refers only to the average score (mean) or the middle score (median) in the norms distribution. It is important to recognize that if the median is used as the norm, 50 percent of the scores must necessarily fall "below the norm."

Local norms developed by school personnel are often more useful than national norms. They are arrived at in the same way as

national norms, but are based on test results obtained locally rather than nationally.

Norms may be expressed in several ways. One way is in the form of grade equivalent scores. A score of 6.4, for example, represents the average score obtained by the norms group in the sixth grade, fourth month. Another and better way is by percentile rank. For example, a score that falls at the 73rd percentile exceeds the scores made by 73 percent of the students in the norms group. Because scores are never precise measures of performance, school personnel are turning with increasing frequency to stanine scores, which represent a range of scores. The distribution is broken into nine segments and a student is ranked from a low range of one to a high of nine, with five being the average range. Intelligence tests yield scores that are expressed as IQs. Originally an IQ score represented the ratio of a student's mental age to his chronological age; recently, however, it has come to represent the difference (deviation) between a student's score and the average scores for students of his own age.

C. SELECTING TESTS

There is no single test that is best for all student populations or for grouping purposes. If a test is to be suitable for a school system, the school personnel responsible for selecting the test should first study the characteristics of the school system and its testing needs. Then they should review the capabilities of the many tests available. They should match the norm groups, reliability samples, and validity studies for each test to the specific population they intend to test. They should review the content of each test to see how closely it matches the curriculum taught in that particular system. Finally, they should examine the scoring system and interpretative materials to determine whether the test is designed for the purpose for which the system intends to use it. Frequently school personnel select an inappropriate

test by failing to go through this selection process and by relying mainly on the word of a test salesman or the comments of a superintendent of an adjacent school system.

First, it is important that school population characteristics be reviewed because tests are peaked for certain levels. Achievement test scores of high school students are all highly related to geographical region, percent of graduates going to college, size of community, and percent of parents who are high school graduates. A science test, for example, may be too hard for a school's seventh graders in a small, rural system, or a test battery geared for the grades 4-6 level may be inappropriate because many children cannot read at that level.

Second, the content of the test must approximately match the curriculum being taught. It is wisest first to read the test publisher's own statement of what the test is designed to measure. Then reference should be made to Buros' *Mental Measurements Yearbooks*,[1] where each test is evaluated by a test or subject matter specialist. Finally, a specimen set of materials should be secured and each test examined item by item to ascertain whether or not it measures the content covered by the teachers. Too often a test is selected because it is packaged nicely and its title sounds particularly good.

Third, the test or battery of tests selected must be determined by the purpose of the testing. While the ultimate purpose of testing is to improve instruction, there are at least five intermediate purposes: placement of students, diagnosis of student weaknesses, assessment of teaching methods, prediction of future student performance, and school evaluation.

An example of testing for student placement would be the use of test scores in the assignment of students to different math levels because the learning of later skills requires the mastery of earlier

[1] Oscar Buros polices the testing industry through securing critics to evaluate tests being published. Each school system should have a set of his *Mental Measurements Yearbooks*. They are so important that no school system can afford to be without them.

ones. In student placement testing, a student's achievement test score is compared with the scores of other students in his school. The prime consideration for the test is that it be reliable enough so that score differences actually represent differences in achievement from one student to another.

An example of testing for diagnosis would be the use of test scores of students who are behind grade level in a particular subject to identify specific skill deficiencies. A student's score on one subtest would be compared with his scores on other subtests. If the testing is being done for diagnostic reasons, the test should have special features. Each subtest, not just the test as a whole, should be reliable. The test should be peaked at a low level because this testing is generally focused on those students who are behind. It should be fairly long since, for example, it requires many more items to analyze weaknesses in each skill than a placement test might require to designate overall achievement in the subject.

An example of testing for assessment of teaching methods would be to compare scores of students taught by a teacher with those of students taught by classroom television or comparing results from students grouped selectively with those grouped heterogeneously. In testing for assessment, a group's scores may be compared with its scores on previous testings. A major consideration in such tests is the equality of the score units at different parts of the score scale with provisions for spanning grade levels.

An example of testing for prediction would be to use test scores to predict a student's chances for success in each course that he wants to select for the following year. In testing for prediction, the student's score on one achievement test may be compared with his score on another. The concern should be with choices among alternative tests and the trustworthiness of differences between scores.

An example of testing for evaluation would be the use of test scores by the superintendent, school board, or state department of education to compare a particular school or school system with

another school or system. In testing for evaluation, a school's achievement test score average may be compared with that of a comparable group of schools. The main consideration is that the content of the test cover the learnings the school or system deems most important.

D. ADMINISTERING TESTS

The way a test is administered can be an important factor affecting test results. Factors like adverse emotional reactions to tests, the race of the examiner, room temperature, and room or corridor noise may have some influence on scores.

There are several common-sense guidelines that school personnel should follow in the administration of tests. They should announce to students well in advance of the testing that a test administration is scheduled and should reassure them that the purpose of the testing is to improve the instructional program. The test administration should not be scheduled before or after a school social or athletic event. Students should be reassured that no one is expected to complete the test in the time allowed. They should also be told whether it is all right to guess or whether wrong answers will be subtracted from right ones. The testing examiner should not convey to the students any personal adverse attitudes toward the testing process. A negative attitude on the part of the examiner can affect the attitude on the part of students. Group testing in an auditorium or cafeteria should be avoided. There should be no interruptions from the outside while the test administration is in progress. The instructions in the examiner's manual should be followed exactly, especially with regard to the time limits and the prohibition against interpreting questions that are unclear to students.

E. INTERPRETING TEST RESULTS

Perhaps the biggest problem with testing is the misinterpretation of test results. Misinterpretation is usually based on false assumptions about what tests can do.

A first false assumption is that IQ tests measure inborn ability to determine for a lifetime a person's potential for learning. What IQ tests measure is a person's ability to perform certain kinds of mental tasks. They measure this performance not at birth but a long time afterwards. The tasks are the kinds that the individual learns as a result of his experiences at home, in school, and elsewhere. An IQ score cannot bypass all the experiences that help or hinder an individual's learning.

Aptitude or IQ tests supposedly differ from achievement tests because they tend to reflect the amount learned from incidental experiences before special training is received whereas achievement tests tend to reflect the amount learned in school. However, in both tests, the abilities tested actually are products of the person's inherited potential for learning and his opportunities for learning. The main difference between them is that the aptitude or IQ test measures tasks learned over a longer period of time than those tasks measured on an achievement test.

The idea that aptitude or IQ tests measure inborn ability has led to the persistent demand that tests be "culture free" or "culture fair." There can be no such thing as a test that is "culture free"; and only a small chance that it can be "culture fair." Some tests, particularly the older ones, are unfair to culturally disadvantaged and minority students. But even those tests that purport to be fair may show lower scores for culturally disadvantaged students because they reflect the racial discrimination, social circumstances, and poor educational opportunities that many of these students have experienced. If opportunities and experience are not equal, the results will not be.

A second false assumption is that a test score is quite reliable, that is, that the score made by a child today will be the

same as the one he will make tomorrow or next week. A test, however, is only a single sample of a child's performance and it can not give more than an estimate of how much he knows.

In school testing situations, the test score may be affected by a number of things. It may be slightly affected, for example, by his mood that day, by whether there is adequate ventilation in the room, or even by whether he has had breakfast or not. It may be affected even more by the questions asked. No test can cover the entire universe of knowledge in any field. A test covers only a sampling of this knowledge. While one student may know the information on which a certain question is based, another student who is just as well informed may not. A student, for example, may know that there are two pints in a quart, but he may not know that there are four quarts in a gallon. This is all to say that tests are not exact; they have a standard error of measurement, and the score obtained is very seldom the true measure of a child's achievement.

The standard error of measurement can be illustrated by reference to one of the most reliable IQ tests ever developed, which has a standard error of 5 IQ points. If a student obtains an IQ score of 100, then there are two chances out of three that the true IQ score falls somewhere between 95 and 105. But there is one chance in three that it does not. There are 95 chances in 100 that it falls somewhere between 90 and 110. And we can be almost certain that it falls somewhere between 85 and 115. So what we have is a possible range from "dull normal" to "above average." This is the degree of accuracy we have on one of the most reliable tests ever devised! Like IQ tests, achievement tests, too, have standard errors of measurement, and many times these are fairly large ones, considerably larger than the standard error mentioned above. Thus, knowing something about the standard error of measurement should make school officials hesitate to label a student on the basis of a single administration of any kind of test.

A third false assumption is that standardized achievement tests measure all that a student knows about each subject tested. As was previously stated, standardized tests do not measure knowl-

edge of the entire subject, but only samples of it. Examination of the questions asked on any standardized achievement test will show that they cover some aspects of the subject that may never have been covered in class and omit things that teachers have emphasized.

A fourth false assumption is that a student's scores on a battery of achievement tests give sufficient information to make decisions about how much a student has accomplished and how well he will be able to do in the future in these subjects. No test battery currently published can do this. There are many important outcomes of learning that cannot be measured by any test battery -- personality, motivation, and creativity -- outcomes that can only be evaluated by teachers and parents who are able to observe the child over a long period of time.

A fifth false assumption is that the profile of scores on the subtests in the subject areas covered by an achievement test battery will reliably show the strengths and weaknesses of the student. For several reasons, the profile does not necessarily do this: the scores may not be the true scores; the score scales may not be comparable; the subtests may have been normed on different student populations; and the scores may not be independent measures but rather highly correlated measures. In looking at an overall profile, a school administrator may remark how well the school has done in reading because the sixth-grade class is reading at an 8.2 level but how poorly the school has done in mathematics because it is achieving at only the 6.1 level. Progress in reading, however, is continuous inside and outside the classroom, whereas progress in arithmetic depends almost entirely upon what is taught in the classroom.

A sixth false assumption is that grade equivalent scores (e.g. 5.2 -- fifth grade, second month) on standardized achievement tests give an accurate picture of the level of a student's performance. Students do not progress at an even pace throughout the school year. Also, over the summer they lose some of what they have previously gained. It may be early November before they are doing work of the kind they were doing the preceding May. Grade equiva-

lents, however, do not reflect the sharply irregular rates of learning or the loss over the summer. Grade equivalents are also misleading because they imply, for example, that a student in the third grade scoring 5.5 on an arithmetic test could be transferred to a fifth-grade class; of course he could not be transferred because he has not learned the skills taught in the fourth grade that he needs for doing fifth-grade work. A much more meaningful description of his score would be that it has a percentile rank of 94, that is, that his score is higher than the scores of 94 percent of the third graders in the norms group.

A seventh false assumption is that a norm is a standard, that it represents what a group of students should be achieving at a particular time. A norm merely describes the performance of students who took the test in the standardization program. Teachers should think carefully about how their students compare in ability with students in the norms group before they decide whether or not they are happy with their test results. National norms should be used simply as reference points. More useful evaluation of test results can be obtained by comparing the performance of a class with that of other classes in the same system through the development of local norms.

An eighth false assumption is that, in spite of knowing the shortcomings of aptitude and achievement tests, school officials can use scores on these tests as the major bases for ability-grouping students by class within school. This is a reprehensible action since such grouping cannot help but result in the segregation of middle-class white children from black children and culturally disadvantaged children on the basis of rather questionable information. It is not that such tests should not be used with underprivileged groups. Rather, until black and poor white families match middle-class whites in jobs, income, housing, quality of education, respect, and complete participation in the day to day affairs of the community, test scores that admittedly depend so heavily on family background and school environment should not be used exclusively in grouping students across the whole school program.

It should be noted here that grouping on the basis of norm-referenced tests will usually result in a different kind of grouping than if the grouping is done on the basis of criterion-referenced tests. With the former, the grouping will be done in terms of how well the student achieves national standards; with the latter, it will be done in terms of how well the student achieves mastery of what is expected for his age and grade. Whereas grouping done with norm-referenced tests may cover entire subjects or entire classes, either permanently or for a substantial length of time, grouping with criterion-referenced tests may be more informal, less permanent, and concerned with segments of subject matter in which certain students have deficiencies that must be overcome before the students can work satisfactorily with the subject matter of the grade.

HOW TO FIND OUT ABOUT ABILITY GROUPING AND TRACKING IN YOUR SCHOOL SYSTEM

A. FRAMEWORK OF STANDARDS

In examining the testing and student grouping practices of your school system, it is desirable to have an overall framework of standards upon which to make a judgment. The following standards adapted from U.S. Department of Health, Education, and Welfare (HEW) regulations[1] are exemplary:

1. Nondiscriminatory tests that are relevant to the purpose of the grouping are used.
2. There is nondiscriminatory application of the standards used for placement.
3. The grouping is for a period in the day only as long as is necessary for accomplishing the purposes for which the grouping is intended.
4. The grouping is designed to meet the special needs of the students in each group by having specially developed curricula or specially trained teachers and staff especially in the slow groups.
5. The system retests frequently enough to determine whether students could be advanced to another group.
6. The system has valid statistical evidence that shows that such grouping is better for the students, including the students who are behind, than other methods of teaching.

If your school system does not meet these standards, then its

[1] The regulations referred to are those of the Emergency School Aid Act (ESAA) C.F.R.§ 185.43. ESAA is a federal program authorizing grants to help school systems meet special needs incident to desegregation.

grouping practices are suspect and may be deemed to be educationally detrimental to some of the students.

To make a judgment about how well your system meets these standards, it is necessary, first, to gather some initial information about the system and then to interview various school personnel like the curriculum director, school counselors, and teachers or principals for the details.

B. INFORMATION INITIALLY NEEDED

It is necessary to become familiar with the locations of schools and of the poor neighborhoods and black neighborhoods. School locations usually can be secured from local chamber of commerce maps or from a master map available at the superintendent's office.

The racial composition of each school can sometimes be secured by a telephone call to the principal. If your system was desegregated by court order, the court order usually requires that the system periodically submit reports to the court listing racial composition by school. These reports can be reviewed at the offices of the civil rights attorneys who worked on the desegregation case or at the federal district court clerk's office if you cannot get cooperation from the school system office. If your school system was desegregated under a voluntary plan approved by HEW, then the racial composition, not only of the schools, but also of selected grades by classroom, can be obtained from the Office for Civil Rights of HEW. Each year, at least through 1975, school systems have submitted to HEW what are called 101 and 102 forms. Not only systems under HEW-approved plans but also systems under court order desegregation plans must submit these forms annually to HEW. These forms may be purchased from HEW for ten cents each, one form per school, by writing and requesting them under the Freedom of

Information Act.[1] Although the local system keeps a copy of these forms, HEW regulations do not require the superintendent to make his copies available for you to see.

The best source for finding in which schools children from low-income families are concentrated, although there is in it no breakdown by race or socioeconomic level, is the system's annual Title I application, which under law is available to any person for public inspection at the local system office.[2]

The reason that gathering all this information is so important is that ability grouping and tracking practices tend to affect black children and poor children more adversely than other children.

C. INTERVIEWING THE CURRICULUM DIRECTOR

In most school systems there is a curriculum director or assistant superintendent or system psychometrist, who is primarily responsible for curriculum development, testing, and grouping of pupils system-wide. You should telephone and ask for an appointment, explaining that you are a member of a citizens' group concerned with quality education.

You will want to put him at ease by approaching him in a manner so as to indicate that you are only interested in learning about the program, not in being critical of it. Begin by asking him to outline his job. Ask him to explain the curriculum and

---

[1] See page 36 for the address of your HEW regional office. The Freedom of Information Act, 5 U.S.C. § 552, guarantees that any person has the right of access to and can receive copies of any document -- subject to nine specific exemptions like top secret information, trade secrets, and personnel files -- so long as the document is in possession of a federal agency.

[2] Title I of the Elementary and Secondary Education Act of 1965, 20 U.S.C. § 241, is a federal program of grants of money to local school systems to be used to meet the special education needs of educationally deprived children in low-income areas. Title I regulations require that the annual applications be shown to any citizen who makes a request. 45 C.F.R. § 116.35.

testing program in general. Then ask for a copy of any written material explaining the curriculum and testing program and any accompanying policies. Try to get a copy of the written material before asking any hard questions that might make him want to change his mind about sharing information with you. The following questions are suggestive of the kinds of things you will want to inquire about, the first few questions being mainly to gain his confidence:

1. What is the size of the school system in terms of student enrollment, number of schools, and budget?
2. What innovations and features of the school program are you especially proud of?
3. How many Title I (poor) children are there in the system? Where are they mostly concentrated?
4. According to test scores, at what level do students perform upon entering and then upon completing their twelve years of schooling?
5. To what elements in the instructional program do you attribute the success of the students?

Elementary Level

6. How are the children grouped at the elementary level?
7. How and when did this grouping evolve?
8. What criteria are used for student grouping?
9. What tests are used? (Try to secure a copy of the publisher's manual for each test.)
10. How were the tests chosen?
11. Do all schools use the same student grouping practices?
12. How are new students arriving in midyear assigned to classes or sections?
13. How do the course content and the method of instruction in the slower groups differ from those in the faster groups?

14. How do the qualifications of the teachers for the slower groups differ from those of the teachers of the faster groups?

15. For what portion of the day or for which subjects does grouping occur?

16. Is the grouping within a single classroom or in several different classrooms?

17. Are children placed in different groups for different subjects on the basis of subtest scores?

18. How are the tests that are used relevant to the purposes of the grouping?

19. If grades are also used for grouping purposes, which grades are used for which subjects and how far back do you go in looking at grades?

20. If recommendations of teachers or counselors are also used, on what criteria do they base their recommendations?

21. How is the cut-off point on scores for placing children in different groups determined?

22. To what extent do students move from one group to another?

23. Has any research been done to validate the local grouping process?

24. Has the progress of slower children been greater than it was before this system of grouping began?

25. Has there been any control group of students who are not ability-grouped with which a comparison of progress has been made?

High School Level

26. To what extent do courses taken at the seventh and eighth grade level limit what can be taken at the ninth grade level?

27. Is the curriculum at the high school level organized for basic, general, and honors levels?

28. At what grade do students begin choosing their own course of study?

29. Besides student preference, what other factors influence what curriculum a student takes?

30. What written information do students and parents receive describing the courses? (Secure a copy of the school system's student handbook.)

31. Is parental permission sought before assignment of students?

32. What are the differences between the college preparatory, general, and vocational curricula? (Secure course descriptions and textbook titles.)

33. Are the full range of honors subjects and subjects in the college preparatory curriculum offered at all high schools?

34. In the subject of English, is the course content different at basic, general, and honors levels or do students cover the same material but at a different pace?

35. Does a student's permanent record or his diploma reflect which course of study he has taken?

36. Do you see any separation of students resulting from the different curricula or the grouping within required high school subjects so that poor or minority students are disproportionately in the lower or slower groups?

37. Are minority and new teachers assigned to teach honors or college preparatory courses?

D. INTERVIEWING COUNSELORS

Try to interview at least one elementary school counselor and one high school counselor, neither of whom is timid and each of whom works at a school where there is a significant number of black students or students from low-income families.

Before asking any difficult questions, first secure copies of all written materials the counselor has to offer, such as counseling rules and the school's student handbook. Begin by asking the counselor to outline the duties that come with his job. Suggestions for areas to explore include the following:

1. What influence do counselors have in assisting in the proper placement of students?

Elementary School Level

2. What is the process used in assigning children to different groups?
3. If tests are used, what is the name, level, and date of publication of each?
4. When, how often, and to whom is each test administered?
5. How was it decided which tests would be used?
6. Who administers the tests?
7. If grades are also used in student placement, what grades are used and for what subjects?
8. If teacher recommendations are also used in student placement, on what are the recommendations based?
9. Is parental permission sought prior to testing?
10. What happens if a parent refuses to abide by the results?
11. What are the cut-off scores for different groups and how are they determined?
12. Is the course content for required subjects about the same for the various groups and the difference in instruction primarily a matter of the pace at which the material is covered?
13. Do individuals teaching the slow groups have the same qualifications as those teaching average and advanced groups?
14. How frequent is retesting?

15. Is there a relationship between test scores and home background?

16. Does grouping put a disproportionate number of black students or students from low-income families in the slower groups?

High School Level

Ask the questions above and questions 26-37 listed on pp. 24-25. Then ask:

17. What colleges recruit at the high school?

18. How have you been involved in determining whether students are college material?

19. How do you counsel students who are interested in going to college but are not college material?

20. By what criteria, other than need, are you guided in selecting and assisting students seeking financial aid?

E. INTERVIEWING PRINCIPALS AND TEACHERS

It would be a good idea to limit your interviews to principals and teachers at schools where there are concentrations of black students or students from low-income families because these are the schools in which any deficiencies in ability grouping and tracking become most evident. Teachers and principals may be asked the same kinds of questions asked of the curriculum director and school counselors. In addition, you may want to inquire as to whether the morale of teachers assigned to the slow groups is lower than that of teachers of other groups, whether new teachers are likely to be given these assignments, and in what proportion minority teachers are assigned there.

Finally, you may want to learn whether, at the high school level, poor and black students are unevenly distributed in sections in required courses such as English, American history, and physical

education. In the case of electives, such as foreign languages, advanced courses, and specialties such as black studies, there may be an uneven distribution simply as a matter of student choice.

You may want to conclude each interview by asking the individual what he personally views as the benefits and drawbacks of such grouping and what changes, if any, he would recommend to remedy the drawbacks.

F. OTHER SOURCES OF INFORMATION

The records of the school system's Title I program have been mentioned previously as a possible source for information. Title I of the Elementary and Secondary Education Act is a federal program of annual grants of money to school systems to be used to meet the special educational needs of educationally deprived children in schools serving concentrations of low-income families. It is the largest federal aid program for school systems. In order to receive Title I funds, each system must annually submit to the state Title I office a Title I application, which under law is available for public inspection at the local school system office.[1] The Title I application tells how the system proposes to spend money to help children who are behind in schools serving low-income areas. To determine which children are educationally deprived, the system will usually test its students. Federal regulations require that there be an evaluation to determine if the program as implemented has helped the children. This is done by pretesting and posttesting the children to measure their gains. Sometimes by reading the narrative section of the Title I applications you may be able to learn what tests were given, whether the children were grouped by test scores, what the deficiencies were, what goals were set, whether they were met, and whether the ability grouping done has worked out well for the children who were behind. Sometimes the goals are

[1] 45 C.F.R. § 116.85. (1975)

set so low that efforts are programmed in such a way that the slower students get further and further behind. And if test results document an expanding gap between the children in the slow group and the children in other groups, then the slow groups are actually being hurt by ability grouping.

Another source of information is the result of the Family Educational Rights and Privacy Act.[1] Under this federal law, any person has a right as a parent not only to see, but also to secure, photostatic copies of any school records directly relating to his child. These records include cumulative folders, discipline reports, grade reports, standardized test scores, psychological evaluations, and written teacher comments. Student records are usually kept in the office of the principal or that of the guidance counselor. Parents have a right to question school personnel concerning what is in the records and to have the officials help interpret or explain the records. Often, if several parents exercise their right to inspect the cumulative record folders of their children, they can determine whether their children have been placed in slow groups according to test scores. This may lead them to ask the school counselor or teacher to explain why children are grouped in a particular way, how the cutoff point was determined, and whether the tests were standardized on schools or on children like their own so as to be an adequate indicator of what the tests are being used to measure.

Finally, an excellent way to obtain information is simply to volunteer to be an aide at a school. Most schools are eager to get parents' help and, during the course of doing volunteer work, parents can observe firsthand the testing procedures and grouping practices.

---

[1] 20 U.S.C. § 1232g (1974).

HOW TO BEGIN THE MOVE FOR REFORM IN YOUR SYSTEM

A. THE "POWER STRUCTURE"

In order to foster reform in your school system, if reform is needed, you must identify the people you have to convince and know what to say to them that will effect change.

School systems, like other institutions, have a power structure. It is usually made up of a board of education, a school board attorney, a superintendent, a number of administrators, the aggregate of teachers, and various community groups. The key person in the power structure is usually the superintendent. However, if ability grouping or tracking is a school practice rather than a systemwide practice, then the principal is a very important person.

The school board is composed of a number of lay persons, either appointed or elected, who are usually paid a nominal salary. They are more likely to respond to an issue than to initiate change on their own. Since they may have neither the time, manpower, nor expertise to handle a particular problem, they will normally refer an issue in which they have an interest to the superintendent. When the superintendent makes a response by way of his administrators, the board normally ratifies or modifies, but occasionally rejects, the superintendent's recommendations. On many issues, boards splinter into different factions. While they have the ultimate legal responsibility for decision-making, they should be the last group, not the first, to challenge. If, on the other hand, there are one or two sympathetic board members, then they should be brought in at the very beginning.

B. PREPARATION AND ORGANIZATION

Before approaching the superintendent, it is desirable to anticipate what advice his administrators will give him when he is confronted with your presentation. For that reason, it is best to talk first with the administrators with whom he will be consulting. More often than not, these are persons you have already interviewed and the very persons who set up the grouping system that caused the problem in the first place.

If during your initial interview you did not ask them about their personal opinions of alternatives to present practices, it might be appropriate to do that now. Some possible alternatives to ability grouping include individualized instruction, heterogeneous grouping, stratified heterogeneous grouping, team teaching, and student tutoring. These are explained in Appendix B.

A second step in preparation for your meeting with the superintendent is to make sure that you have organized a coalition of parents. One inquiring parent is soon forgotten. Try to secure the cooperation of prominent persons in the community whom top school officials respect. Solicit the support or endorsement of the local teacher organization, if there is one, by tying the issue to such matters as the amount of classroom time occupied by testing or the possible use of the tests by administrators for teacher accountability. Enlist the help of one or two PTA presidents to bolster your support. Perhaps there are community organizations with which you have not previously dealt but which might provide valuable support. Black and white coalitions carry considerable clout.

If you do not have an organization, create one. Sometimes a group can be formed simply by getting together several parents who want to visit the school to examine the school's cumulative records of their children. Alternatively, a group may be formed by recruiting people who would be interested in being placed on a Title I

Parent Advisory Council at their school.[1] Or you may piggy-back your concerns with a separate group that is forming around another issue such as getting rid of an incompetent principal or remedying unfair student suspension practices.

There should be someone in your organization who is familiar with the ins and outs of your school system and the personalities involved. Recruit for chairperson an individual who has some free time and a good public presence. Have the group help you compile a report or adopt the report that you have prepared. Obtain access to a mimeograph machine to reproduce your report, making the format as attractive as possible. It is particularly important that each person in your group be conversant and articulate in all aspects of the issue. In order to create a solid and consistent information base for your organization, it may be advisable to hold one or more private sessions for the purpose of self-education. This enables each of your members to be a confident spokesperson for your viewpoint and creates an image of collective credibility. Once you have become established, your organization can best widen its visibility and educate the public to its concerns through the press and through contact with other community groups, such as the League of Women Voters, PTAs, church groups, and civic clubs. Exposure to convictions differing from yours, coupled with the experience of speaking to groups about the issue, are good ways of preparing for your meetings with the superintendent and, later, with the school board.

---

[1] Title I regulations require that there be formed a Parent Advisory Council at each school receiving Title I funds. 45 C.F.R. § 116a.25(a). They meet several times a year and are supposed to exert considerable influence over how money is to be spent to meet the educational needs of eligible children. 45 C.F.R. § 116.25(g). Each school Parent Advisory Council elects one or more persons to represent it on a systemwide Parent Advisory Council, which advises the school system in the planning, implementation, and evaluation of the total Title I program. 45 C.F.R. §§ 116a.25 (a),(d), and (e).

C. MEETING WITH THE SUPERINTENDENT

To meet with the superintendent, call in advance to secure an appointment. A delegation of about five people from your group or groups endorsing your position should plan to attend the meeting. Parents with children who are directly affected should make up most of the delegation. The discussion, however, should not center around individual grievances but rather around systemwide practices. Shortly before arriving at the meeting, the delegation may wish to rehearse who is going to say what. The most effective presentation is simple and brief. Each item should be separately numbered rather than combined into a nebulous statement. You may want to pause after each item to allow the superintendent to give his reaction. If his opinions or promises are vague or adverse, at the end of the meeting request that he formalize his response to each item in writing and secure from him an approximate date by which the response will be ready. More often than not, superintendents' responses are simply to explain and justify rather than to promise to reconsider or to change. For that reason, it almost always becomes then necessary to take your case before the school board.

D. MEETING WITH THE SCHOOL BOARD

Unless the meeting with the superintendent achieves the goals you seek, the meeting with the school board should be scheduled as soon as you know that the board will see you. If your visit is to be at the time of the regular business meeting, call a couple of weeks early in order to have your issue placed on the agenda near the beginning of the meeting. Items of business taken up toward the end of such meetings tend to receive less attention and are dealt with more summarily.

A large turnout of the members of your group and other groups supporting you is essential to show your strength. The presentation itself should be a formal written statement that is succinct and clear. It should contain the normal courtesies extended to

public officials. Copies of it should be distributed to the board members and the news media present just prior to the presentation. The presentation should be made by the most effective spokesperson in your group. It should identify the groups that endorse your position, the previous discussions with administrators and other groups, and the need for the board to take immediate action in reviewing and reconsidering the present ability-grouping practices. School board members should be requested to make individual comments on the statement.

E. PRESS RELATIONS AND FOLLOW-UP

A single appearance before the school board most likely will not be sufficient to develop a move for reform. That is why it is so important to generate some rapport with the news media. Notify the news media in advance of your community meetings when you expect a respectable turnout. When you plan for different things or when events come up, issue periodic press statements and, where the occasion warrants such a step, hold press conferences. For example, you might hold a press conference to announce the formation of a coalition, another after meeting with the superintendent, and another after the superintendent's written response is received. After you make a presentation to the school board, your contact with individual reporters should become personal and ongoing.

The presentation to the school board should be followed up with visits to each board member to discuss the matter further in private. Continued visibility is essential through talks at meetings of PTAs and community groups, guest appearances on local radio or television talk shows, and subsequent news articles or letters to the editor in the local paper.

Because no change will occur unless there is recognition of the problem, occasionally citizen groups must use direct action to dramatize the issue. Such actions may include petition drives, personal protests by withdrawal of children during the days of

testing, raising at annual school board budget hearings the issue of whether too much money is being spent on testing, or even calling for a boycott of the local business of a hostile school board member.

There are a number of things that impede the possibility of change in the school system. In addition to the school administration's own belief in the perceived educational benefits of ability grouping, there may be a desire to accede to the wishes of more vocal middle-class parents who want their children to be in honors or college preparatory groups, away from slow or culturally deprived children. Also, most teachers may prefer to teach homogeneous groups because it makes their job easier by limiting their class groups to students learning at the same level and at the same pace, and putting students who are discipline problems as well as slow learners in classes taught by someone else. In addition, some school administrators, unfortunately, may still be of the belief that black children are less capable than whites and that ability grouping will keep blacks away from whites so that whites will not be pulled down. In school systems with heavy black enrollments, white school officials may fear that abandoning ability grouping might cause white flight to private or parochial schools. Finally, sales representatives of some test publishers have been able to convince some school administrators that for every educational ill there is a test that will remedy it, and these school administrators may have constructed an elaborate testing program that is easier to allow to perpetuate itself than to dismantle.

If results are lacking and interest begins to wane, then it may be desirable to seek outside help.

F. OUTSIDE HELP

Reform may be more meaningful and lasting if accomplished without outside help for often local initiative is destroyed when such help is brought in. Use of outside help, however, becomes necessary when local efforts are not getting anywhere. Following are the outside

resources that you should know about:

1. Education Section Chief
   Civil Rights Division
   U.S. Department of Justice
   Washington, D.C. 20530
   (202) 739-4092

2. Regional Director
   Office for Civil Rights
   Department of Health,
   Education, and Welfare

50 Seventh Street, N.E. -- (Alabama, Florida, Georgia
Atlanta, Georgia 30323 Mississippi, North Carolina,
(404) 526-3312 South Carolina, Tennessee)

1114 Commerce Street -- (Arkansas, Louisiana, and
Dallas, Texas 75202 Texas)
(214) 749-3301

Gateway Building -- (Virginia)
3535 Market Street
Philadelphia, Pennsylvania 19104
(215) 596-6772

3. Director-Counsel
   NAACP Legal Defense and
   Educational Fund, Inc.
   10 Columbus Circle, Suite 2030
   New York, New York
   (212) 586-8397

4. Executive Director
   Southern Regional Council
   52 Fairlie Street, N.W.
   Atlanta, Georgia 30303
   (404) 522-8764

If your school system has undergone desegregation without a federal court order, then you should contact HEW. HEW will investigate ability grouping practices to determine whether they violate Title VI of the 1964 Civil Rights Act, a law prohibiting racial discrimination in any program receiving federal funds, or the Emergency School Aid Act (ESAA), a law providing federal funds to school

systems to meet the special needs incident to desegregation, but making systems ineligible for this money if they maintain any practice, such as ability grouping, that racially isolates or otherwise discriminates against students.

If your system has undergone desegregation under a federal court order, then jurisdiction would lie with the Justice Department. But since Justice has responded in a very slow and conservative manner to challenges to ability grouping, it is also desirable to contact the NAACP Legal Defense Fund, which often is more responsive. One of their local cooperating attorneys may have handled the desegregation case in the past or may be willing to become involved now.

Even though your system desegregated under a federal court order, you may nevertheless be able to bring HEW in. First, if your system happens to be applying for or is receiving ESAA funds from HEW, then HEW may investigate to determine whether there is noncompliance with ESAA regulations forbidding ability-grouping and tracking practices causing racially identifiable classes.[1] Secondly, even though your system is under court order and has not applied for ESAA funds from HEW, HEW has some authority to become involved. An order of a federal court in Washington, D.C., requires HEW "to monitor all school districts under court desegregation orders to the extent that their resources permit and to bring their findings to the attention of the court concerned."[2] To date, however, HEW unfortunately has not exercised this authority on the grounds that its manpower resources will not allow such monitoring.

Regardless of whether you seek the assistance of the Justice

---

[1] 45 C.F.R. § 185.43(c). A racially identifiable class is one in which the racial composition of the class varies more than ± 20 percent from the racial composition of the grade. For example, if grade 4 at Smithville Elementary School is 50 percent black, then, to be racially identifiable, one or more classes in that grade would have to be less than 30 percent black or more than 70 percent black.

[2] Adams v. Richardson, 356 F. Supp. 92, 99 (D.D.C. 1973), mod. and aff'd. 480 F.2d 1159 (D.C. Cir. 1973), 391 F. Supp. 269 (D.D.C. 1975).

Department, the Legal Defense Fund, or HEW, you should thoroughly document your complaint, explaining not only the results of your group's investigations but also its attempts to get the school system to remedy the situation.

The Southern Regional Council can provide limited help by assisting you in drafting complaints, following through with federal agencies, and helping you to secure cooperating Legal Defense Fund attorneys. The Council may also be able to put you in touch with resource organizations in your state such as the Delta Ministry, the American Friends Service Committee, and others who have long-established expertise in education issues, especially civil rights problems.

ABILITY GROUPING AND THE LAW

There has been no court decision directly declaring the practice of ability grouping or tracking students to be unconstitutional per se. Rather, court decisions on ability grouping have always turned on whether the grouping has caused classrooms to become racially identifiable and, if so, whether this is based on test scores which reflect the continuing effects of past racial discrimination.

In the past, courts were reluctant to become involved in the issue of ability grouping because they felt it was a matter best left to educators.[1]

However, in 1967, in a landmark case, a federal district court judge found the Washington, D.C., track system, as administered, to be unconstitutional because it deprived poor and black students of their right to an equal educational opportunity.[2] The court found that the schools in the poorest Washington neighborhoods had the highest proportions of students in the lower tracks and that the lowest tracks had the highest percentages of black students in them. The court found that students were placed into tracks on the basis of standardized tests inappropriate for poor and black student populations, that each track offered a different kind of education both in pace and in scope of subject matter, that there was very little movement of students between the tracks, and that the low track had a watered-down curriculum instead of a compensatory program to ameliorate specific academic deficits. The whole scheme was found fatally defective because student placement was not based upon capacity to learn but upon intelligence tests standardized on a middle-class white population. As a result, placement was based

---

[1] Miller v. School District No. 2, Clarendon Co., S.C., 256 F.Supp. 370 (D.S.C. 1966); Steel v. Savannah-Chatham County Board of Education, 333 F.2d 55, 61-2 (5th Cir. 1964).

[2] Hobson v. Hansen, 269 F.Supp. 401 (D.D.C. 1967) aff'd. sub. nom., Smuck v. Hobson, 408 F.2d 175 (D.C. Cir. 1969).

more on socioeconomic status and race than on ability and most black students were wrongly placed in low tracks, depriving them of an equal educational opportunity.

By 1970 most of the school systems in the South were completing the process of desegregating their schools. Some of them decided that placing black students in the same classrooms with their white counterparts would lower the quality of education because of the inferior education provided to black students under the former dual systems. This caused some school systems to assign students to rooms according to achievement test scores. Courts, however, rejected this approach because it tended to create segregated classes within school buildings. Rather than delving into the issue of the validity of testing, the courts simply ruled that testing could not be employed to group students or assign them to schools until after the school systems had already been completely desegregated.[1] Thus, ability grouping could not be used as a component of a desegregation plan.

In some systems, once the desegregation process occurred, there were immediate attempts thereafter to assign students on the basis of ability grouping. In these instances, the courts also declared that such a practice was constitutionally impermissible.[2] The courts stated that there would have to be a minimum waiting period of several years before grouping could be done.[3] The clearest explanation for this waiting period was stated in a case in which the court said that assigning black students in a recently desegregated school on the basis of scores on standardized ability and achievement tests violated their rights to be treated equally with white students when the black students had recently been

---

[1] Singleton v. Jackson Municipal Sep. School District, 419 F.2d 1211 (5th Cir. 1970); U.S. v. Tunica County School District, 421 F.2d 1236 (5th Cir. 1970); U.S. v. Sunflower County School District, 430 F.2d 839 (5th Cir. 1970).

[2] U.S. v. Humphreys County School District, Civil Action No. GC 6645-S (N.D. Miss. 1971).

[3] Lemon v. Bossier Parish School Board, 444 F.2d 1400 (5th Cir. 1971).

educated in inferior schools and, in taking tests, were competing with white students educated in superior schools.[1]

Finally, in 1975, the U.S. Fifth Circuit Court of Appeals declared that ability grouping resulting in racially segregated classrooms could not be used by any school system that had previously been segregated even if such grouping had been used as an educational technique years before desegregation.[2] School administrators, the court declared, must operate completely integrated systems without ability grouping for a period long enough to assure that the underachievement of the slower groups was not due to the educational disparities caused by prior segregation. How long is enough has not been stated, but, presumably, if a student had attended a segregated school for, say, six years, it could be argued that he could not legally be ability-grouped for another six years or however long it might take to bring him up to peer status and end the educational disadvantages accorded him by prior segregation.

One of the most difficult problems concerning any constitutional attack on the practice of ability grouping in the post-desegregation era is that of proof. In a recent case, for example, a court held that an inference of discrimination could not be drawn from facts showing that in a school system with three grouping levels, over 47 percent of the white students but only 20 percent of the nonwhites were in the high level and 39 percent of the nonwhite students but only 11 percent of the whites were in the low group.[3]

In order to mount a successful court challenge of ability grouping, significant evidence must be gathered to prove that (a) there is a high correlation between race and student assignment, (b) each level offers a different kind of education both in pace

---

[1] Moses v. Washington Parish School Board, 330F. Supp. 1340 (E.D, La, 1971).

[2] McNeal v. Tate County School District, 508 F.2d 1017 (5th Cir. 1975).

[3] Morales v. Shannon, 516 F.2d 411 (5th Cir. 1975).

and scope of subject matter with the lowest being the worst, (c) there is very little mobility between the different levels, (d) there is no adequate compensatory program in the lower level sufficient to help students overcome the effects of past segregation or present cultural disadvantages, and (e) test scores are more likely a reflection of the past education of a student or his present socioeconomic background than a reflection of his capacity to learn.

Presently pending litigation challenging the constitutionality of the ability grouping schemes in various school systems demonstrates the size and complexity of the task of proving a case to the satisfaction of a federal court.[1] Expert witnesses or consultants may have to be hired to testify about the selection of inappropriate tests by the school system and about erroneous conclusions made by school personnel in interpreting the scores or using them in student placement. Experts may have to direct the gathering of selected data from the school system concerning the effect of ability grouping on different groups of students, interpret these data to the lawyers and to the court, and, if called upon to do so, may have to offer educationally sound alternatives to the practices being challenged.

Abuses in ability grouping have been challenged also by HEW in administrative proceedings. Under the Emergency School Aid Act (ESAA)[2] (a program authorizing federal grants to help school districts meet special needs incident to desegregation) no school system is eligible to receive such funds if it assigns students in a way that creates classroom segregation. Regulations implementing the law provide that where there are racially identifiable classes, a presumption is created that the method of student assignment is racially discriminatory and the burden shifts to school officials to demonstrate that the grouping is bona fide ability grouping used

---

[1] See, e.g., Scott v. Winston-Salem/Forsyth County Board of Education, Civil Action No. C-174-WS-68 (M.D.N.C. 1973)

[2] 20 U.S.C. § 1603 (1974).

as a standard pedagogical practice.[1] The criteria for determining whether the grouping is bona fide are whether it is (1) based on nondiscriminatory, objective standards of measurement which are educationally relevant to the purposes of such grouping; (2) determined by the nondiscriminatory application of standards and maintained for only as long in the school day as is necessary; (3) designed to meet the special needs of students in each group and to improve student achievement in lower groups by specially developed curricula, specially trained personnel, and periodic retesting for promotion; and (4) validated by test scores or other reliable objective evidence indicating the educational benefits of such grouping.[2] These regulations have been upheld by the courts.[3]

To determine whether ability grouping in a school system is beneficial or detrimental to students, HEW compares the achievement gain or loss for students in the low group with the achievement gain or loss for students in other sections -- middle, high, etc. -- of the same grade at the same school. If, for example, the slow group at a particular school ranked at the thirtieth percentile when compared with the national norms at the third grade and ranked at the fifteenth percentile when tested at the fifth grade, while the middle group at the same school ranked at the fiftieth percentile in both third and fifth grades, then the grouping has not resulted in equal educational benefit for the two groups.

Few, if any, Southern school systems with racially identifiable classes have been able to sustain the burden of showing that their grouping is bona fide. Although less than one fourth of the Southern systems having racial minorities apply for ESAA funding and fewer than ten percent receive such funds, the regulations have had some impact. Under this program, HEW has stated that it has caused the reassignment of pupils out of racially identifiable classrooms

---

[1] 45 C.F.R. § 185.43(c).

[2] Ibid.

[3] Board of Education of Cincinnati v. Department of H.E.W., 396 F.Supp. 203 (1975).

in about 140 Southern school systems to date. However, many of the school systems misusing ability grouping have declined to apply for the funds to avoid becoming entangled in civil rights compliance problems in the first place.

In school systems which do apply for or receive ESAA funds, the regulations are very helpful, because parents or citizens in these systems can write HEW asking the federal agency to investigate any systemwide abuses. If HEW does decide to investigate, then the school system can be required to furnish detailed information not readily available to citizen groups and the bulk of the technical work of analysis is borne by HEW equal opportunity specialists. The federal agency findings may form a basis upon which subsequent litigation may be initiated should HEW be unsuccessful in remedying the problem with the local system. HEW findings in the form of letters of ineligibility or noncompliance, and often backup documents, can be secured from HEW under the Freedom of Information Act.

APPENDIX A

PREVALENCE OF ABILITY GROUPING ACROSS THE SOUTH

How frequent is the utilization of ability grouping across the South and to what extent does it result in racially identifiable classes in Southern schools? These are two questions the Southern Regional Council sought to answer by reviewing data submitted to the Department of Health, Education, and Welfare (HEW) by local school systems during the 1973-74 school year.

Data on the school systems of seven of the eleven Old Confederate States are available at the Region IV Office for Civil Rights of HEW in Atlanta. These seven states -- Alabama, Florida, Georgia, Mississippi, North Carolina, South Carolina, and Tennessee -- have approximately 937 school systems and enroll about 6.5 million students, of whom about 2 million or 31 percent are black, about 84,000 or 1 percent are Spanish American, and 18,000 or .25 percent are Native American.

Each year school systems receiving federal funds from HEW must submit to the agency what are called 101 and 102 forms. The forms are filled out by the local systems themselves and list such information as districtwide and school-by-school enrollment figures by race and special figures on racial composition of classes of sample grades.

The data for the 1973-74 school year were reviewed to determine all schools in each system that local administrators marked as utilizing any form of ability grouping of students as part of the regular school program. The data were reviewed, in addition, to determine all schools in each system where pupil assignment within the school caused racially identifiable classes. A racially identifiable class is one in which the racial composition of the class varies more than ± 20 percent from the racial composition of the grade at that school. For example, if grade 4 at Smithville Elementary School is 50 percent black, then to be racially identifiable,

one or more classes in that grade would have to be less than 30 percent black or more than 70 percent black.

Several general conclusions can be drawn from the data reviewed. First, across the region roughly two out of every three districts have racially identifiable classes in one or more schools. Secondly, seven out of every ten districts themselves state that they use ability grouping as a means of placing students. Of those districts that do use ability grouping, seven out of every ten also have racially identifiable classrooms. With the exception of Mississippi, less than one out of every ten districts having schools with racially identifiable classes do not use ability grouping.

More specifically, here is a descriptive statistical breakdown by state:

ALABAMA

Of the 124 systems in Alabama, 68 or 55% have racially identifiable classrooms in one or more schools. Ninety-three or 75% of all districts claim to use ability grouping. Sixty-three percent of those 93 have racially identifiable classrooms in one or more schools. Only 13% of the 68 districts with racially identifiable classrooms do not use ability grouping.

FLORIDA

Of the 67 systems in Florida, 58 or 87% have racially identifiable classrooms in one or more schools. Fifty-five or 82% of the systems claim to use ability grouping. Ninety-six percent of those 55 districts have racially identifiable classrooms in one or more schools. Only 9% of the 58 districts with racially identifiable classes do not use ability grouping.

GEORGIA

Of the 188 systems in Georgia, 134 or 71% have racially identifiable classrooms in one or more schools. One hundred forty-three or 76% of the systems claim to use ability grouping. Eighty-three percent of those 143 districts using ability grouping have racially identifiable classrooms in one or more schools. Only 11% of the 134 districts with racially identifiable classes do not use ability grouping.

MISSISSIPPI

Of the 158 systems in the state, 90 or 57% have racially identifiable classrooms in one or more schools. One hundred five or 66% of the systems use ability grouping. The percentage of those 105 systems using ability grouping and having racially identifiable classrooms in one or more schools could not be determined. The percentage of those 90 systems with racially identifiable classrooms not using ability grouping could not be determined.

NORTH CAROLINA

Of the 159 systems in the state, 99 or 62% have racially identifiable classrooms in one or more schools. Ninety-nine or 62% of the systems use ability grouping. Ninety-three percent of those 99 systems using ability grouping have racially identifiable classrooms in one or more schools. Only 7% of the 99 districts with racially identifiable classes do not use ability grouping.

SOUTH CAROLINA

Of the 94 systems in the state, 80 or 85% have racially identifiable classrooms in one or more schools. Seventy-two or 77% of the systems use ability grouping. One hundred percent of those 72 systems using ability grouping have racially identifiable classrooms in one or more schools. Only 10% of the 80 districts with racially identifiable classes do not use ability grouping.

TENNESSEE

Of the 147 systems in the state, only 71 were surveyed. OCR has not continued to require data from districts with no minority student enrollment. Of those 71, 39 or 55% have racially identifiable classrooms in one or more schools. Forty-six or 69% use ability grouping. Eighty-three percent of those 46 systems using ability grouping have racially identifiable classrooms in one or more schools. Only 7% of the 39 districts with racially identifiable classes do not use ability grouping.

In short, there seems to be a definite relationship between the prevalence of ability grouping and the prevalence of racially identifiable classrooms in Southern schools.

There are certain limitations, however, that must be placed on this conclusion. First, the data do not address themselves to

the correlations between schools that ability-group and schools where racially identifiable classrooms exist. Rather the statistical descriptions show only correlations between systems with these characteristics. It is often the case that the practice of ability grouping is not districtwide and often the case that the existence of racially identifiable classrooms is not systemwide. In a rare instance, within a single system, one or more of the schools may use ability grouping without having racially identifiable classrooms and other schools may have racially identifiable classrooms but no ability grouping.

Secondly, ability grouping at the high school level resulting in racially identifiable classrooms may be explained in some cases because instruction is departmentalized and student interest in course selection may occasionally split along racial lines for certain electives. Only in the required subjects like English, American history, and physical education can it be determined whether such grouping is racial. Unfortunately, the 1973-74 HEW data do not single out required courses from electives and it is sometimes not possible to distinguish a legitimate natural grouping situation from an illegitimately engineered one.

REFERENCE

Mills, Roger. "Testing, Ability Grouping, and School Desegregation," Paper presented at the Southeastern Invitational Conference on Measurement in Education, University of Tennessee, Knoxville, Tennessee, December 7, 1974.

APPENDIX B

ERIC ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

Conducted by Educational Testing Service in Association with Rutgers University Graduate School of Education

TM REPORT 3 JUNE 1971

ABILITY GROUPING: STATUS, IMPACT, AND ALTERNATIVES

Miriam M. Bryan

How widespread is the use of ability grouping in the public schools of the United States? To what extent do tests represent an integral feature of ability grouping plans? What are the effects of ability grouping on the scholastic achievement and on the personal and social development of students so grouped? Is ability grouping likely to result in ethnic and socioeconomic separation within the school? What have test publishers done to determine and/or ensure the usefulness and the fairness of their tests for students who are culturally different? What have researchers reported concerning the reliability and validity of tests they have used with the disadvantaged student? What are some of the alternative strategies to ability grouping that have proved to be effective in the improvement of instruction?

These questions were among many to which answers were sought by a group of specialists in educational measurement commissioned in late 1969 by the U.S. Office of Education to study the status of ability grouping in American public schools and its impact upon the academic and affective development of school children (Findley and Bryan, 1971).

Ability grouping, as defined in that study, is "the practice of organizing *classroom groups* in a graded school to put together children of a given age and grade who have most nearly the same standing on measures or judgments of learning achievement or capability." Grouping and regrouping *within a classroom* for instruction in particular subjects is not considered to be ability grouping in the sense of this definition.

Ability grouping has been a topic of debate for more than half a century. The issue has, however, been brought into sharper focus during the last several years by three developments: (1) the launching of Sputnik and the consequent emphasis on special education for students with superior capabilities to meet the need for highly trained scientists; (2) increased attention to special education for the mentally and physically handicapped; and (3) emerging concern for equality of educational opportunity for all children, with obvious implications for the improvement and enhancement of that opportunity for those children to whom it has previously been denied.

In spite of the admission that homogeneous grouping by ability across the subjects of the school curriculum is impossible and in spite of conflicting evidence gathered over the years as to the benefits of ability grouping, such grouping is widely practiced in the nation's public schools. While grouping occurs in school districts of all sizes, it is especially characteristic of larger school systems; and while done at all grade levels, it is more common in the higher grades than in the lower grades. There is proportionately more grouping in the Northeast and Middle West than in other parts of the country.

While a relatively small proportion of schools rely on test scores alone for ability grouping, virtually all ability grouping plans depend on tests of aptitude and/or achievement as an integral feature. Findley and Bryan (1971) found that test scores alone constituted the basis for grouping in 13 per cent of the school districts reporting, but were among the multiple criteria reported by 82 per cent. Other criteria included teacher, counselor, and/or principal judgment, school grades, and student and/or parent interest, or a combination of these.

Although ability grouping is widely approved by school administrators and school teachers, opinion polls show that an overwhelming number of teachers express preference for average, mixed, or superior classroom groups over classes of low ability, in which emotional disturbance and rebellious behavior, as well as poor achievement, are likely to abound. Research on "streaming" (ability grouping) in England's schools indicates that the most detrimental effects occur in "non-streamed" classes taught by "pro-streaming" teachers. This generalization could apply equally well to American schools.

Early research studies on ability grouping were almost entirely concerned with the effect of grouping on academic achievement. While the evidence, then as now, was conflicting, the earlier studies more often than not reported

The Clearinghouse operates under contract with the U.S. Department of Health, Education and Welfare, Office of Education. Contractors are encouraged to express freely their judgment in professional and technical matters. Points of view expressed within do not necessarily, therefore, represent the opinions or policy of any agency of the United States Government.

ABILITY GROUPING: STATUS, IMPACT, AND ALTERNATIVES

Miriam M. Bryan

gains by low groups and losses by high groups when compared with similar students taught in heterogeneous classes. More recent studies tend to show that separation into ability groups, when all children involved are considered, has no clear-cut positive or negative effect on average academic achievement, and the slight trend toward improving achievement in superior groups is counterbalanced by poorer achievement in the average and low groups, particularly the latter. One possible explanation for this difference is that in the earlier period the prevailing emphasis in instruction was on drill, with strong academic motivation accepted as a favorable but not a necessary characteristic, while today both strong academic motivation and academic achievement are emphasized; another is that low-achieving groups contain far more children of minority and low socioeconomic groups today than they did earlier, when the comparisons were between groups within a narrower range of ethnic and socioeconomic variation.

Research evidence regarding the effect of ability grouping on the affective development of students has, until recently, been very thin, perhaps because emotional and social growth is more difficult to assess than intellectual growth. As with the studies of impact on achievement, there has been little uniformity among the findings reported for the research studies that have been made. However, much of the evidence, especially the more recent evidence with ethnic and socioeconomic overtones, supports the generalization that the effect of ability grouping on the affective development of students is to reinforce favorable self-concepts in those assigned to high achievement groups and to reinforce unfavorable self-concepts in those assigned to low achievement groups. Low self-concept operates against motivation for academic achievement in all students, but especially among those from minority groups and lower socioeconomic backgrounds.

Most recently, researchers have become concerned with the effect of ability grouping on ethnic and socioeconomic separation. Here the evidence has been more conclusive. Students from minority groups and from unfavorable socioeconomic backgrounds te d to score lower on tests and to be judged less accomplished by teachers than students from middle-class homes. To the extent that these students are over-represented in low ability groups, then, they are being made to suffer the unfavorable results of ability grouping. A grouping plan which creates classes where disadvantaged students are in the majority deprives them of the stimulation of middle-class children as learning models and helpers, and commonly produces poorer achievement on their part. The greatest positive impact on the school learning of disadvantaged children occurs when the proportion of middle-class children in a group is highest.

Children of many minority groups come disproportionately from lower socioeconomic backgrounds. The disadvantages of their backgrounds are further compounded by language disabilities. For some of them, English, in which teaching and testing are generally done, is a "second language"; for others, the language patterns differ markedly from "standard American English." Language disabilities not only have the direct effect of making learning more difficult, but also have the indirect effect of lowering self-concept because of frequent correction.

There have been no studies to date of the reliability and validity of tests administered to culturally limited populations for the specific purpose of ability grouping. As a matter of fact, until recently few publishers have studied the general usefulness of their tests with disadvantaged students. Now systematic efforts are being made by test publishers and research agencies to review present test offerings and to introduce new emphases to meet the particular problem of assessing the capabilities and achievement of the disadvantaged group.

The research that has been done to date shows that standardized aptitude tests, as they are currently constructed, are no less reliable for disadvantaged students than they are for others. They do, however, tend to overpredict for the disadvantaged group; that is, the disadvantaged student may not perform subsequently as successfully as his tests scores indicate that he should. The same findings, in a slightly more limited way, apply also to standardized achievement tests. This is not to say that certain items in a standardized test may be more easily answered by students of one culture than by those of another, but, rather, that minority students who select the intended responses do not always perform up to expectations. The evidence that tests standardized on other populations tend to overpredict the subsequent performance of disadvantaged students and, hence, are not unfair to them, is less than comforting. The challenge is to develop ways of describing learning progress directly rather than to settle for measures that are "fair" only in the sense that they reflect "fairly" the results of educational disadvantages.

Generally speaking, researchers are not studying or trying out and evaluating tests. They are studying other matters and, with few expections, accept uncritically the standardized test and/or use it as the best available instrument at hand. In the search of the literature concerning the use of tests in ability grouping and, especially, with the use of tests with the culturally deprived, several misuses of tests were noted. Among these, the following should be mentioned: (1) assuming that a test designed for students of a given age or of an estimated ability level can be used indiscriminately with students of different ages and/or experiences; (2) modifying the test in some material respect, but still applying the regular norms (for example, changing items or answers because of local circumstances; or translating the entire test into another language); (3) testing so early in preschool programs that culturally deprived children are not even ready to manipulate the test materials; (4) testing so early in preschool programs that

ABILITY GROUPING:
STATUS, IMPACT, AND ALTERNATIVES

Miriam M. Bryan

there is no opportunity for children with limited backgrounds to "learn" the abilities tested; (5) using tests written in standard American English, with heavy emphasis on vocabulary, for students for whom this is a second language or who speak in a particular dialectic style; (6) testing very small numbers of students over a very short period of time; (7) failing to follow through for two, three, or four years or more; (8) interpreting scores of individual students on short subtests when reliability estimates make it impossible to trust such interpretations; (9) treating different measures of learning ability as though the results on them were comparable; and (10) attaching the same importance to predictive validity without intervention (in the form of compensatory training) as with it.

The research concerned with ability grouping and with the procedures for the use of tests in grouping students for learning has provided only limited information. The design for the research procedures, the selection of tests, and the interpretation of test results have frequently been questionable. Most important, the research has produced inconclusive and conflicting results. This applies equally to the research findings concerning the advantages and disadvantages of ability grouping and to those regarding either the validity of currently available tests for use with culturally limited students or the validity of the interpretations of the test results for such students.

If, then, present ability grouping practices seem inadequate, what alternative strategies are there? The six suggestions that follow do not exhaust all possible alternatives, but they are judged to be the most promising for the promotion of learning:

1. *Individualized instruction*. There are almost as many definitions of individualized instruction as there are "authorities" defining the term. It is thought of here as instruction of the individual student, once his characteristics have been defined, by the prescription of sequences of learning experiences leading to the mastery of basic skills and structural knowledge.

2. *Heterogeneous grouping*. This involves the putting together, in unselective fashion, of students who may vary extensively in age, experience, and knowledge and may, therefore, have opportunities to learn from one another that are not always provided by homogeneous grouping. Heterogeneous grouping of this kind is practiced in the nongraded school.

3. *Stratified heterogeneous grouping*. Grouping of this kind—notably the Baltimore plan of stratified heterogeneous grouping by tens—takes into account the concern for curtailing extreme heterogeneity, while allowing for enough diversity to give leadership opportunities in each class and avoiding the concentration of defeated and stigmatized students in a low group almost impossible to inspire or teach. In the Baltimore plan, 90 students ranked in order of excellence on some composite—a standardized test battery, for example—are then subdivided into nine groups of ten each. Teacher A is given a class consisting of the highest or first 10, the fourth 10, and the seventh 10; Teacher B has the second, fifth, and eighth tens; and Teacher C has the third, the sixth, and the ninth tens. In this kind of grouping there is no top or bottom section; each class has a narrower range than the full 90 students have; teachers can give attention where it is needed without feeling that there are extremes whose needs are not being met; no teacher has to teach a class of disruptive children who lack both motivation and capability.

4. *Team teaching*. Several different models for team teaching have been developed. Each model embraces the concepts of individualized instruction, mastery, and differentiated staff working under the leadership of coordinating master teachers. Students who need to learn the same tasks may work in groups assigned to a designated teacher for the purpose of learning the special tasks. The grouping is informal, ad hoc, and of short duration. Such grouping promotes the effective utilization of the personnel and resources, and increased learning by the individual student, without the detrimental effects of homogeneous grouping.

5. *Student tutoring*. In student tutoring plans, top students within a class may help those having difficulty with various subjects; or older children may be "imported," and perhaps paid, to tutor younger children who are having difficulty in learning the basic skills. Such tutoring works to the advantage of both groups of students. In fact, tutors who were themselves academically retarded have been found to gain even more than the tutored.

6. *Early childhood education*. Such education applies to the provision of opportunities for all children, especially those in need of compensatory education, to enjoy intellectual stimulation in a supportive emotional climate, at least from kindergarten at age five and somewhat earlier when possible. Competence generated by the nature of early stimulation should increase the readiness of the children to participate in the conventional schooling of the primary grades.

Taken together, these alternative strategies constitute a constructive challenge to the uncertain advantages and the harmful effects of ability grouping on academic achievement, affective development, and the ethnic and socioeconomic separation of children. In each of them, tests and other evaluative measures may be used constructively if they are used with care and caution.

ABILITY GROUPING:
STATUS, IMPACT, AND ALTERNATIVES

Miriam M. Bryan

In conclusion, the following recommendations are offered: (1) Ability grouping, as defined, should not be used; however, flexible grouping within classes may be used to advantage when the information gained by testing and/or observation is the first step in a program of diagnosis and individualized instruction. (2) In any grouping plan, provision should be made, as part of the instructional program, for frequent review of each student's grouping status. (3) Alternative strategies for ability grouping should be explored and exploited for their usefulness in promoting learning. (4) Favorable self-concept should be a goal in itself, but it is also a supportive factor in learning. An attitude of firm confidence and hope by the teacher, fundamental to effective learning, should be conveyed to every student. (5) Teacher training should include an emphasis on welcoming diversity in children, especially with regard to language and customs of minority groups, and on teaching children to prize it in each other. (6) Finally, steps should be taken *as early as possible* in each local situation to promote unitary school populations in each district and in each classroom. Action to improve instruction by any of the alternative strategies to ability grouping will be effective in proportion to the extent to which they can be applied before a district or city has become almost completely an ethnic and/or a socio-economically limited population.

REFERENCE

Findley, W.G., & Bryan, M.M. *Ability Grouping: 1970*. Athens, Georgia: Center for Educational Improvement, College of Education, University of Georgia, 1971. (96 pp.)

APPENDIX C

DESCRIPTIONS OF TESTS COMMONLY USED IN ABILITY GROUPING, *with Critical Comments on Each*

Despite the favorable comments on the characteristics of many of many of these tests, no one of them should be used as the sole criterion for general ability grouping.

MENTAL ABILITY (OR INTELLIGENCE) TESTS

Cooperative Preschool Inventory. Educational Testing Service, Revised, 1970.

Age Levels: 3-6 (1 form)

Content: Items measure achievement in areas necessary for success in school: the child's knowledge of his personal world and his ability to follow verbal instructions; his knowledge of time sequences, locational associations, and characteristics of certain social roles; his familiarity with basic numerical concepts, judgments of "more or less," knowledge of positional relationships; and metric shapes, size, speed, weight, and color.

Raw Scores: Total score only

Derived Scores: Percentile ranks at six-month intervals for children age 3-0 to 6-11

Norming: Based on children in 11 Head Start centers that included whites, blacks, Mexican-Americans, Polynesians, and other minority groups; data are reported for those children tested in English. In addition to national norms for children of middle and low socioeconomic status, there are regional norms.

Reliability: Estimated internal consistency reliabilities for various age groups range from .86 to .92, sufficiently high for individual assessment at the ages for which the Inventory is intended. The standard error of measurement for the age groups varies from 3.1 to 3.9.

Validity: Concurrent validity coefficients based on 1,476 subjects taking the Stanford-Binet during the standardization range from .39 at age 3 to .65 at age 5.

• *Comment:* The Inventory is not culture-free. One of its aims is to permit educators to assess the degree of disadvantage a child has on entering school so that indicated deficiencies may be overcome. The Inventory may also be used to demonstrate changes associated with educational experiences. It was originally designed under the sponsorship of the Office of Economic Opportunity for use in the

Head Start program and is recommended for use in this and other preschool programs. The Inventory is individually administered and should be followed up with individualized instruction.

Cooperative School and College Ability Tests (SCAT), Series II. Educational Testing Service, 1967-73.

Grade Levels: 4-14 (4 levels, 2 parallel forms per level)

Raw Scores: 3 scores -- verbal, quantitative, and total

Derived Scores: Converted scores, percentile ranks, and percentile bands

Norming: Standardization was done on a sample made up of a small number of students in a large number of schools representing systems of various sizes and different geographic locations.

Reliability: Internal consistency estimates range from the upper .80s to the middle .90s; however, since the subtests are rather speeded, these estimates may be misleading.

Validity: Validity evidence compares favorably with that reported for other intelligence tests. Average validity coefficients between total score and school grades range from .59 to .68.

• *Comment:* These are good instruments for obtaining gross group measures. They should not be used for individual assessment.

The Henmon-Nelson Tests of Mental Ability. Houghton Mifflin Company, 1973 Revision.

Grade Levels: K-12 (4 levels, 2 parallel forms per level)

Content: Primary Battery -- listening, picture vocabulary, size and number tests. Upper levels -- 90 items of ten different types, all verbal and quantitative in nature, but more verbal than quantitative

Raw Scores: Total score only

Derived Scores: By Age -- deviation IQs (DIQs), stanines, and percentile ranks of DIQs. By Grade -- stanines and percentile ranks of raw scores

Norming: The tests were standardized on a national sample that included communities of different sizes and different socioeconomic levels. While rural, urban, and suburban schools were all represented in the sample, no special information is given as to the extent of the participation of minority groups in the standardization sample. For the primary battery, however, correlation coefficients are reported for first-grade children from both disadvantaged and nondisadvantaged backgrounds.

Reliability: Split-half reliabilities for upper level tests are high (.93 to .96). However, since the tests are timed, these estimates may be spuriously high.

Validity: Content validity evidence is convincing. Criterion-related validity evidence shows correlations between scores on tests at all levels with the various subtests in the Iowa Tests of Basic Skills ranging from .49 to .71 for the primary battery and from .80 to .86 for the upper levels. The primary battery correlates only reasonably high (.57+) with other ability tests designed for this level.

• *Comment:* The primary battery is designed to measure the verbal and quantitative skills important in assessing readiness for school work. The influence of reading skills on test performance is completely eliminated; instructions and questions are presented orally by the examiner, and students respond to questions by marking appropriate pictures or symbols in their test booklets. All levels of the test seem suitable for obtaining group measures. Like all tests of their type, however, their use should not result in permanent grouping.

Lorge-Thorndike Intelligence Tests. Houghton Mifflin Company, 1954-1966.

Grade Levels: K-13 (2 parallel forms for each of 10 levels)

Content: Verbal Battery -- vocabulary, sentence completion, arithmetic reasoning, verbal classification, and verbal analogy. Nonverbal Battery -- pictorial classification, numerical relationships, and pictorial analogy

Raw Scores: Total score on each battery

Derived Scores: Verbal and nonverbal IQs, composite (average of V and NV), verbal age and nonverbal age equivalents, verbal and nonverbal grade equivalents, verbal and nonverbal grade percentiles. All IQs are deviation IQs (DIQs) with a mean of 100 and a standard deviation of 16.

Norming: Based on approximately 19,000 students per grade for grades 3-12 from communities across the United States stratified on size, family income, and median education of adults in the community

Reliability: Alternate forms estimates range from .83 to .94 for the verbal battery and from .80 to .92 for the nonverbal battery. Split-half reliabilities are all above .90. Standard errors of measurement range from 2.4 to 6.1 DIQs. The scores appear to be quite stable, correlations between scores obtained a year apart ranging from .52 to .88, and from .49 to .55 after a three-year difference.

Validity: Descriptive evidence demonstrates high content validity for the tests. Correlations with the verbal battery and the subtests of the Iowa Tests of Basic Skills range from .72 to .84; the nonverbal battery correlations run lower (.57 to .68). Correlations between the Lorge-Thorndike and average grades two years later range from .39 (nonverbal) to .56 (verbal). Correlations with other well-known intelligence tests range from the low .60s to the middle .80s.

• *Comment:* The tests are intended to be measures of abstract intelligence defined as "the ability to use and interpret symbols." Two of the major assets of the norming procedure are that (1) the norms are comparable from grade to grade and (2) the Iowa Tests of Basic Skills and the Tests of Academic Progress were normed on the same school systems, thus providing the opportunity to compare intelligence and achievement scores. The tests should be classified as standing high among the better measures of their kind; the decision to be made by the test user is whether the tests are suitable for use with his particular student population.

Otis-Lennon Mental Ability Test. Harcourt Brace Jovanovich, 1970.

Grade Levels: K-12 (6 levels, 2 parallel forms per level)

Content: Pictorial/geometric classification, pictorial/geometric analogies, following directions, quantitative reasoning, picture vocabulary, and general information

Raw Scores: Total score only

Derived Scores: By Age -- deviation IQs (DIQs), percentile ranks, stanines, and mental age equivalents (for three lower levels). By Grade -- percentile ranks and stanines

Norming: Based on roughly 200,000 students in 117 school systems from all 50 states, selected to be representative of the entire United States educational system. School systems were stratified on the basis of (1) enrollment, (2) public, private, church-related, (3) socioeconomic index, and (4) geographic region.

Reliability: Median alternate forms reliability was .92; median split-half reliability, .95; median internal consistency reliability, .94. Median correlation between scores on the test administered one year apart was .87. Standard errors of measurement in DIQ points averaged about 6.0 for ages 5 to 9 and about 4.3 for ages 10 through 17.

Validity: Content validity evidence is demonstrated by items and item types; criterion-related validity evidence shows correlations generally in the .70s between the Otis-Lennon and achievement tests and school grades; construct validity evidence shows correlations ranging from .70 to .90 between the Otis-Lennon and readiness and differential aptitude tests.

• *Comment:* The authors stress the fact that the tests do not measure innate capacity and that test scores can and do change across time. They clearly state that the assessment of mental ability rests upon the basic assumptions that all students have had equal opportunity to learn the types of things included in the tests and are equally motivated while taking the test. They readily admit that these assumptions may not hold for children who have been severely culturally deprived. They should, therefore, not be used with such children.

Peabody Picture Vocabulary Test. American Guidance Service, 1959-70.

Age Levels: 2.5-18 (2 parallel forms)

Content: 150 test plates, each with 4 numbered pictures; examiner reads a stimulus word and the subject responds by pointing to, giving the number of, or otherwise indicating the picture best illustrating the word.

Raw Scores: Total score only

Derived Scores: Percentile ranks, mental ages, and deviation IQs (DIQs) with a mean of 100 and a standard deviation of 15. Because the publisher uses 6-month (through 5 years) and 12-month chronological age classifications rather than smaller intervals, there are big "jumps" in the IQ table.

Norming: Standardization was based entirely on 4,012 white children and youth in and around Nashville, Tennessee.

Reliability: Alternate form estimates range from .67 at the 6-year-old level to .84 at the 17- and 18-year-old levels.

Validity: Although several validity studies are mentioned in the examiner's manual, the author states that "all of the statistical validity on the test are (sic) limited and preliminary."

• *Comment:* The test is of moderate reliability and largely unestablished validity. Considerable caution needs to be used in interpreting the norms, especially in communities other than Nashville. While the test could probably be used as a quick estimate of intelligence for normal white children, the PPVT is probably the least satisfactory intelligence test among those reviewed for this handbook -- and it has little to recommend it for use with minority groups.

READINESS TESTS

Gates-MacGinitie Reading Tests: Readiness Skills. Teachers College Press, 1969.

Grade Levels: K-1 (1 form)

Raw Scores: 9 scores -- listening comprehension, auditory discrimination, visual discrimination, following directions, letter recognition, visual-auditory coordination, auditory blending, word recognition, and total (all preferably weighted)

Derived Scores: Standard scores and percentile ranks

Norming: No information is given about the normative sample other than to describe it as being nationwide, consisting of "approximately 4500 children in 35 communities. . . carefully selected on the basis of size, geographic location, average educational level, and family income."

Reliability: Reasonably satisfactory reliabilities are reported for the subtests but the reliability of the total score is not, giving the impression that performance on the various subtests is the critical factor rather than performance on the test as a whole. This is contrary to usual professional thinking.

Validity: Correlations between total readiness scores obtained in October and vocabulary and comprehension scores on the Gates-MacGinitie Reading Test taken the following May are reported as .60 and .59, respectively. Presumably, then, the test has good predictive validity; however, no help is given to users in interpreting performance from a predictive point of view.

• *Comment:* This is a very long test (2 hours in administration time, longer than that of any other readiness test); the relative merits of the longer testing time are for the user to decide. It is unfortunate that the authors never do explain exactly what the test is designed to measure nor do they provide completely adequate interpretative data for test results. The test may be useful in grouping children within beginning reading classes; it should not be used for grouping children generally.

Metropolitan Readiness Tests. Harcourt Brace Jovanovich, Inc., 1965-69.

Grade Levels: K-1 (2 parallel forms)

Raw Scores: 7 scores -- word meaning, listening, matching, alphabet (knowledge of lower-case letters), numbers, copying, and total; Draw-a-Man is optional.

Derived Scores: Percentile ranks and stanines. Scores are also expressed in terms of five-level readiness status ratings.

Norming: Although considerable information is provided about the nature of the standardization group, it is not clear how representative this group is of first-grade students as a whole. However, the authors make a strong case for underplaying the importance of national norms for predictive validity, stressing instead the relationship between an obtained readiness score and later achievement.

Reliability: Reliability data, computed by both split-half and alternate form techniques, are generally above .90 for total score; subtest reliabilities range from .50 to .86. The authors downplay the usefulness of subtest scores because of the relatively low reliabilities.

Validity: The authors describe the validity of the test by showing the relevance of the content; by demonstrating the test's relationship with other measures of school readiness like the Murphy-Durrell Reading Readiness Analysis, and by relating success on the MRT with performance on the Stanford Achievement Test: Reading for first-grade students at each of the five readiness levels.

• *Comment:* These tests appear to measure abilities commonly believed to be associated with early school learning. Unusually

specific information is provided concerning the instructional significance of the test results. While the tests should not be used for permanent class grouping, they can be very useful in determining programs for individualized instruction with small groups within the classroom.

Lee-Clark Reading Readiness Test. California Test Bureau, Revised 1962.

Grade Levels: K-1 (1 form)

Raw Scores: 4 scores -- letter symbols, concepts, word symbols, and total

Derived Scores: Percentile ranks and grade equivalent scores

Norming: Based on two different populations (normal and above-normal intelligence), resulting in some problems in score interpretation. Norms for entering first-graders and end-of-year kindergarten students are based on different but unspecified populations.

Reliability: Split-half reliability estimates based on 170 first-graders range from .87 to .96; however, since the subtests are timed, these estimates should be interpreted cautiously.

Validity: Predictive validity estimates are in the .40s and .50s.

• *Comment:* This test should serve as a good screening device and provide a fairly gross measure for initial, but temporary, grouping purposes within reading classes. It is not a valid diagnostic test even though the test authors recommend its use for diagnostic purposes.

ACHIEVEMENT TESTS

California Achievement Tests. California Test Bureau, 1970 Edition.

Grade Levels: 1.5-12 (5 levels, 1 form)

Raw Scores: 11 or 12 scores -- reading (vocabulary, comprehension, total), mathematics (computation, concepts and problems, total), language (auding [lowest level only], mechanics, usage and structure, total), (spelling scored separately from other language subtests), and composite score

Derived Scores: Grade placement scores, percentile ranks, standard scores, stanines

Norming: The standardization was done jointly with the Short Form Test of Academic Aptitude on a sample of approximately 203,684 students. A stratified random sample of school districts represented all 50 states grouped in seven geographic regions; small, medium,

and large districts; and urban, rural, town, and miscellaneous community types. A simple, random sample of schools from within each district was selected; entire grades were tested in these schools. Catholic schools were sampled separately.

Reliability: Numerous reliability coefficients are reported for the various subtests at the various levels. Internal consistency reliabilities for total subtest scores at the various levels range from .85 to .96. Reliabilities estimated for the total scores for all five levels are consistently high, either .96 or .97.

Validity: Content validity was stressed in the construction of all subtests. Construct validity was studied by obtaining correlations between the subtests of the California Achievement Tests and those of other achievement batteries. While the correlations are not reported, they are described as substantial.

• *Comment:* The authors have attempted to develop a battery of achievement-diagnostic tests. However, using scores on these tests or their subtests for diagnostic purposes is not recommended until more data on the reliability of test score differences is available. Subtests in reading, arithmetic, and language should be useful for temporary grouping within the specific areas. Under no circumstances should the composite score be used for grouping by classes.

Comprehensive Tests of Basic Skills, Expanded Edition (CTBS/S). California Test Bureau, 1968-1973.

Grade Levels: K-13 (2 forms, 7 overlapping levels)

Raw Scores: 12 to 15 depending upon level -- reading (vocabulary, comprehension, total), language (mechanics, expression, spelling, total), arithmetic (computation, concepts, applications, total), and battery total. In addition to the three basic skills which are common to all levels, batteries at each level from grades 1.5 to 12.9 contain subtests in science and social studies; batteries at several levels from grades 2.5-12.9 cover reference skills as well; scores on these tests are not included in the total battery score.

Derived Scores: Percentile ranks, stanines, and standard scores for all levels; in addition, grade equivalents are reported for all grade levels above 1.5. For several levels from grade 2.5 to grade 12.9 Anticipated Grade Equivalents and Anticipated Achievement Scale Scores are reported.

Norming: Norms are based on a large national sample of students in grades K-12 and include students in both private and public schools.

Reliability: Internal consistency estimates for total score range from .98 to .99. Information concerning the reliability of the subtests at the various levels is not yet available.

Validity: Content validity was stressed in the test construction. To help eliminate cultural bias, approximately 20 percent of

the pretest samples were blacks. It would be helpful to have some data about concurrent and predictive validity.

• *Comment:* Commendable features of the CTBS/S include (1) removing items that might have racial or ethnic bias, (2) having the examiner read questions in several subtests at various levels aloud to reduce the effects of reading ability, (3) providing a practice test at the lower levels, and (4) emphasizing the higher mental processes rather than the measurement of factual knowledge per se. The tests are recommended for temporary within-classroom grouping in reading, language, and arithmetic. Under no circumstances should the score on the total battery be used for any kind of grouping.

Iowa Tests of Basic Skills. Multilevel Edition (Forms 5 and 6). Houghton Mifflin Company, 1971.

Grade Level: 3-8/9 (2 forms)

Raw Scores: 15 scores -- vocabulary, reading comprehension, language skills (spelling, capitalization, punctuation, usage, total), work-study skills (map reading, reading graphs and tables, knowledge and use of reference materials, total), mathematic skills (mathematics concepts, mathematics problem solving, total), and composite score

Derived Scores: Grade equivalents, age equivalents, and standard scores, with their own percentile ranks and stanines. Grade equivalent conversion tables separately for each level and subtest, and average grade equivalents for total language, total work-study skills, total mathematics, and composite score are provided.

Norming: The standardization was very carefully done, the norms group closely representing the national school population generally. Norms are provided for individual students and for school averages. National percentile norms for grade equivalents are given for the beginning, middle, and end of school year, thus making it unnecessary for schools using the battery to restrict themselves to a single testing period or to depend on extrapolation of between-testing norms. Special grade equivalent norms are also provided for regional areas, Catholic schools, and large city groups. In addition, national percentile norms for age equivalents, for age groups, and for standard scores are also available.

Reliability: Of the 84 reliability coefficients reported on the subtests, only 6 are in the .70s; the others are in the .80s and .90s. The composite score reliabilities are all .98. Standard errors of measurement are reported for each grade for raw scores and grade equivalents.

Validity: Content validity was emphasized in the test construction, and the thoroughness with which it was done is a major strength of the battery. Test users will find interpretive materials related to the content very beneficial for planning remedial instruction.

• *Comment:* This is a thorough battery of tests designed to "pro-

vide for comprehensive and continuous measurement of growth in skills that are crucial to current day-to-day learning activities as well as to future educational development." Subtest scores should be useful for grouping within specific subject areas. The composite score should not be used for grouping of any kind.

NOTE: Until 1972 the Iowa Tests of Basic Skills were designed for use in grades 3-8/9. Now a primary battery has been designed for use in grades 1.7-3.5. This battery includes 15 subtests in listening, vocabulary, word analysis, reading comprehension, language skills (spelling, capitalization, punctuation, usage), work-study skills (maps/graphs and tables, references), and mathematics skills (mathematics concepts, mathematics problems). It yields 15 subtest scores and a composite score. In all characteristics, it resembles the multilevel edition intended for upper grades.

Iowa Tests of Educational Development (ITED). Revised Edition. Science Research Associates, Inc., 1963.

Grade Levels: 9-12 (2 forms)

Raw Scores: 11 subtest scores, no total score -- understanding of basic social concepts, general background in natural sciences, correctness and appropriateness of expression, ability to do quantitative thinking, interpretation of literary materials, general vocabulary, and uses of sources of information

Derived Scores: Standard scores and percentile ranks

Norming: Based on scores of over 50,000 students in 136 school systems in 39 states tested in the fall of 1962. The population of the schools was stratified according to geographical location and school size.

Reliability: The split-half reliabilities of the full-length version of the tests range from .82 to .95, and from .83 to .96 for the classroom version (all tests except use of sources of information). Composite score reliabilities range as high as .99. The probable error of any single standard score is approximately 1.2 points.

Validity: The authors discuss content, predictive, construct, and concurrent validity, but unfortunately they do little more than discuss them. Claims for other than content validity are not substantiated either by convincing data or in sufficient detail.

• *Comment:* From a technical standpoint, the ITED appear to have been well constructed. A valuable feature is the inclusion of profiles for high school students who have already graduated from college, with separate profiles provided for each major field for students with A, B, and C averages. Expectancy tables have also been developed so that one can predict scores on the tests in the American College Testing Program and the College Entrance Examination Tests as well as probable success in college. The ITED were designed, as the title indicates, as broad measures of general educational development. They were not intended to be used for grouping purposes and should not be so used.

Metropolitan Achievement Tests (MAT). Harcourt Brace Jovanovich, Inc., 1970 Edition.

Grade Levels: K-9.5 (6 levels, 2 forms)

Content: Primer (K-1.4) -- listening for sounds, reading, numbers. Primary 1 (1.5-2.4) -- reading (word knowledge, reading), word analysis, mathematics (concepts, computation). Primary 2 (2.5-3.4) -- reading (word knowledge, reading), word analysis, spelling, mathematics (computation, concepts, problem solving). Elementary (3.5-4.9) -- reading (word knowledge, reading), language, spelling, mathematics (computation, concepts, problem solving). Intermediate (5.0-6.9) -- same as for elementary level, plus science, social studies. Advanced (7.9-9.5) -- same as for intermediate level

Raw Scores: 3 to 9 subtest scores, plus total reading and total mathematics scores, depending upon level; no composite score

Derived Scores: Standard scores, percentile ranks, stanines, grade equivalents

Norming: The major variables used in selecting and describing the standardization group were (1) socioeconomic index -- based on median family income and median years of schooling of persons over age 24 in the sample communities, (2) size of community, (3) geographic region, (4) public vs. nonpublic system, (5) mental ability test scores. The characteristics of the Metropolitan standardization samples and the national population are highly comparable. The battery was standardized at two separate times during the school year, fall and spring. Approximately 7,000 students per grade took each form of the test except for the Primer level, where the samples were approximately 1,500 students per grade. All students in both the fall and spring programs took the Otis-Lennon Mental Ability Test in the fall so that directly comparable data would be available for the two groups. As a result of an elaborate standardization program, the standard scores provided for all forms and all levels are on a common scale; percentile ranks and norms are provided for end of kindergarten, middle of grade 1, beginning and end of grades 2 through 8, and beginning of grade 9; stanines and stanine norms are provided for the beginning and end of each grade; and grade equivalents are reported on a scale that extends from 1.0 to 9.9.

Reliability: Estimates of internal consistency reliability (Kuder-Richardson 20) range from .85 for mathematics concepts at the Primary 2 level to .96 for at least nine subtests of various types and subtest totals at levels beyond Primer; split-half reliability estimates range from .88 for mathematics computation at the intermediate level to .97 for seven subtest and total subtest scores from the Primary 1 battery on. Median Kuder-Richardson 21 estimates of reliability range from .78 for mathematics computation for grade 5 to .96 for total reading for grade 3. The standard errors of measurement range from 1.7 to 4.1 raw scores.

Validity: The authors of the battery discuss validity chiefly in terms of content validity, which, of course, is most important in any critical evaluation of an achievement battery. With regard to the appropriateness of the tests for use at the local level, the

authors rightly take the point of view that this is a matter to be determined locally.

• *Comment:* Without question, the Metropolitan Achievement Tests, 1970 Edition, constitute one of the finest achievement batteries available. The scores on the various subtests can be useful in temporary grouping within the classroom; the fact that no composite score is provided is a strong deterrent to the use of this battery in ability grouping generally.

Sequential Tests of Educational Progress, Series II. Educational Testing Service, 1971.

Grade Levels: 4-14 (4 levels, 2 forms)

Raw Scores: 7 scores for grades 4-12, 5 scores for grades 13-14, no total score -- English expression, reading, mechanics of writing (spelling, capitalization, and punctuation), mathematics computation, mathematics basic concepts, science, social studies. The writing and mathematics computation tests are available for grades 4-12 only.

Derived Scores: Converted scores, percentile ranks, percentile bands, stanines

Norming: The major portion of the standardization program was conducted in the spring, when the final forms were administered to a nationwide sample of about 106,000 students in grades 3 through 12 and in grade 14. The college level tests had been administered the previous fall to a nationwide sample of approximately 1,400 students in grade 13 to obtain data for entering college freshmen. The norms samples were quite closely representative of numbers and percentages of public elementary and secondary school students, of Catholic students, and of college freshmen in the national population. Both individual and school mean norms are provided.

Reliability: Two types of reliability estimates are provided: internal consistency coefficients, computed using Kuder-Richardson Formula 20, and parallel forms product-moment correlations. Of the 200 internal consistency coefficients for the two forms of the tests, 82 are .90 or higher; 93 are in the .85 to .89 range; and 25 are below .85. For both forms, the greater number of lower reliabilities are at grade 3, with mathematics basic concepts and spelling showing up most poorly. In general, the size of the parallel forms correlation coefficients indicates that the parallel forms of each test measure essentially similar competencies. Standard errors of measurement for the individual test forms are relatively low.

Validity: Since STEP Series II was designed to assess developed abilities in seven broad areas of education, the content validity of the tests is of major importance. Test construction procedures set up to ensure such validity were apparently effective.

• *Comment:* STEP Series II, combined with the Cooperative Primary Tests, provides continuous measurement from grade 1 through grade 14. While this is a good battery, each prospective test user

should, before a decision is made to use it, examine the content of each test in the battery in order to evaluate content validity with respect to his own instructional practices. The battery is intended for use in measuring group achievement, class, grade, school, or system. *It is not intended for use for grouping purposes.*

Achievement Series: SRA Achievement Survey. Science Research Associates, 1975.

Grade Levels: 1-9 (5 levels -- two forms); Primary Edition (1-2, 2-4 -- two overlapping levels); Multilevel Edition (4-5, 6-7, 8-9 -- three overlapping levels): Blue Level, Grades 3-5; Green Level, Grades 5-7; Red Level, Grades 7-9.

Content: Primary Edition -- reading, language arts, mathematics. Multilevel Edition -- reading (comprehension, vocabulary), language arts (usage, spelling), mathematics (concepts, computation), social studies (optional), science (optional), use of sources optional)

Raw Scores: Primary Edition -- 3 subtest scores and composite score, plus 28 (Primary 1) or 30 (Primary 2) optional skill scores. Multilevel Edition -- 9 subtest and subtest total scores, plus composite score, or 13 scores if optional tests are used, plus 40 optional skill scores.

Derived Scores: Standard scores, grade equivalents, percentile ranks, special percentiles, stanines, deciles, growth scale values

Norming: Approximately 156,000 students from 6,500 classrooms in 816 schools in 220 school districts from nine geographical regions and eight classes of districts participated in the standardization program. Numbers of students by grade ranged from 7,281 in kindergarten to 17,880 at grade 5. In addition to national norms, separate subgroup norms were set up for Title I students, large city schools, nonpublic schools, high socioeconomic schools, and town/rural schools.

Reliability: Kuder-Richardson Formula 20 estimates of internal consistency for subtests of the Primary Edition range from .92 to .94, with most of the estimates in the high .80s or low .90s; composite score estimates range from .95 to .97; for the Multilevel Edition, estimates range from .79 to .94 for the separate subtests, from .87 to .96 for subtest totals, and from .96 to .98 for composite scores. Standard errors of measurement are given in both raw score units and for derived scales.

Validity: The content of the battery was based on studies of elementary school curricula, basal text series, and supplementary teaching materials to identify objectives common to the various programs and to determine the grade level at which they were taught. Curriculum specialists and SRA content experts suggested source materials, recommended item writers, and provided assistance in the writing of objectives.

• *Comment:* SRA Assessment Survey is the title used by the publisher for the combination of the batteries described above and the Iowa Tests of Educational Development to cover the grade range 1-12. A feature of the SRA Assessment Survey is the provision of growth scales, numeric scales, one for each subject-matter area, that cover student performance and provide continuous measurement in growth scale values from grades 1-12, extremely useful to schools or school districts interested in making longitudinal studies of progress in learning. As with the other batteries reviewed here, subtest scores may be helpful in temporary grouping within the classroom; under no circumstances should composite scores be used for grouping by classes.

Stanford Achievement Test. Harcourt Brace Jovanovich, Inc., 1973 Revision.

Grade Levels: 1.5-13 (8 levels, 6 complete batteries for Primary through Advanced levels, and partial batteries for TASK I and II (Test of Academic Skills); 2 forms of complete batteries, 3 forms of TASK)

Content: Primary 1 (1.5-2.4) -- vocabulary, reading comprehension, word study skills, mathematics concepts, mathematics computation, spelling, listening comprehension. Primary 2 (2.5-3.4) -- same as Primary 1, plus mathematics applications, social science, science. Primary 3 (3.5-4.4), Intermediate I (4.5-5.4.), Intermediate II (5.5-6.9) -- same as Primary 2, plus language. Advanced (7-9.5) -- same as Intermediate II, excluding word study skills, listening comprehension. TASK I (9-10) and TASK II (11-13) -- reading comprehension, mathematics concepts, language

Raw Scores: All levels except TASK report subtest scores, total reading, total mathematics, and total battery scores; all levels except Advanced and TASK report a total auditory score; TASK reports reading, mathematics, and language scores, with no total score.

Derived Scores: Grade equivalents, percentile ranks, stanines, and scaled scores; there are no norms for the scaled scores.

Norming: For the Primary through the Advanced levels, a stratified sample (109 school systems drawn from 43 states) provided over 275,000 students for three standardization programs (October and May programs to reflect beginning-of-year and end-of-year performance and, for Primary 1 and 2, a February program to reflect midyear performance). The appropriate level of the Otis-Lennon Ability Test was also administered to provide test users with a means whereby the achievement test score might be compared with intelligence. For TASK separate norms for high school students by grade (8-12) and junior/community college freshmen are based on an October testing of 46,491 students in grades 8-13 from 29 states.

Reliability: Reliability data are in the form of split-half reliability coefficients, Kuder-Richardson estimates of internal consistency, and standard errors of measurement. The 75 split-half coefficients range from .67 to .96, with all but six being above .85. Standard errors of measurement range from 2.0 to 4.0 raw score

points. Of the 60 split-half and Kuder-Richardson coefficients reported for each form and both levels of TASK, all are above .92. The 30 standard errors of measurement range from 2.5 to 3.4.

Validity: Both content and construct validity were stressed during test development. A major goal of the authors was to make sure that the test content would be in harmony with present-day school objectives and would measure what is actually taught in today's school. There is no evidence given of predictive validity.

• *Comment:* A major advantage of this battery is that it provides for a continuous measurement of skills, knowledge, and understanding in basic school subjects from grades 1.5 through 13. While the Stanford is no doubt one of the most carefully constructed tests with respect to reflecting the curriculum in our public schools, test users must be cautious in their use of the Stanford as the sole criterion of what should be taught in their classes. As with other good achievement batteries, scores on subtests may be useful in temporary grouping within the classroom; students should not, however, be grouped by classes on the basis of the total battery scores provided.

Tests of Academic Progress (TAP). Houghton Mifflin Company, 1964-72.

Grade Levels: 9-12 (1 level, 3 forms)

Raw Scores: 7 scores -- social studies, composition, science, reading, mathematics, literature, and total

Derived Scores: Standard scores and percentile ranks

Norming: Normative data for the TAP were obtained from a coordinated standardization program that also involved the Iowa Tests of Basic Skills and the Lorge Thorndike Intelligence Tests. The sample used in establishing norms was selected so as to be representative of public and parochial school students in the nation. Percentile norms are provided for each grade for fall, midyear, and spring testing. In addition, norms for school averages are available.

Reliability: Split-half reliability coefficients are mostly in the high .80s or low .90s. No test-retest reliabilities are given; it is likely that such coefficients would have been somewhat lower.

Validity: The manuals for this test give little information on validity. However, for each subtest, the teacher's manual provides rather extensive breakdown of the content covered by the items.

• *Comment:* The authors have been successful in developing items that test a variety of cognitive abilities, with emphasis on such higher order abilities as interpretation, comprehension, evaluation, and application of principles and procedures rather than on recall of information only. The authors have been wise to focus their testing on six "basic skill" content areas, in which they have attempted to base items largely on the abilities that would be developed in relatively basic courses and to do little with the content of advanced courses. If TAP is used with care and judgment, these instruments could prove very valuable in counseling and in assessing the academic progress of a secondary school. TAP should not, how-

ever, any more than any of the other batteries reviewed, be used in permanent ability grouping; the subtest scores should be useful for grouping within the classroom, but under no circumstances should the composite score be used for grouping by class.

Wide-Range Achievement Test. Revised Edition. Guidance Associates of Delaware, Inc., 1940-65.

Age Levels: 5.0-11.11, 12.0 and over (2 levels, 1 form)

Raw Scores: 3 subtest scores, no total score -- reading (recognizing and pronouncing words), spelling (copying marks resembling letters, writing the name, writing single words to dictation), arithmetic (counting, reading number symbols, solving oral problems, performing written computations)

Derived Scores: Grade equivalents, standard scores, and percentile ranks

Norming: The authors report that no attempt was made to obtain a representative national sampling for norming purposes. No data are provided concerning the samples but a brief comment is made that an attempt was made to use IQs available from a variety of tests to develop norms "that would correspond to the achievement of mentally average groups with representative dispersions of scores above and below the mean."

Reliability: The authors report questionably high split-half reliability coefficients by one-year age groups. A low coefficient of .981 out of 14 coefficients on the reading test makes all the coefficients suspect. Furthermore, all parts of the test are timed, if not speeded, a feature that would cause one to discount the startlingly high values that the authors report

Validity: The authors cite 11 ways in which the test has reportedly been found of value as "an adjunct to tests of intelligence and behavior adjustment"; however, no statistical evidence or research studies are reported to support this claim.

• *Comment:* This "achievement" test is a unique, individually administered test. Careful examination of the materials leads one to seriously question why the authors chose to label this an "achievement" test. While the test might be a useful clinical tool for a psychologist working with specialized cases, it is impractical for general school use and should be discouraged for such.

READING TESTS

Gates-MacGinitie Reading Tests. Teachers College Press, 1965-1972.

Grade Levels: 1-12 (7 levels, including one overlapping test for grades 2.5-3; 2 or 3 forms depending upon level)

Content: Primary A, B, C (grades 1, 2, 3) -- vocabulary and comprehension. Primary CS (grades 2.5-3) -- speed and accuracy. Survey D. E. F (Grades 4-12) -- speed and accuracy, vocabulary, comprehension.

Raw Scores: Primary A, B, C -- vocabulary, comprehension, and total. Primary CS -- speed and accuracy. Survey D, E, F -- speed and accuracy, vocabulary, comprehension, and total. Tables are given to interpret gain scores for either individual students or the whole class.

Derived Scores: Grade equivalent scores, standard scores, and percentile ranks

Norming: Separate norms correspond to different testing periods used in the standardization -- October and May for Primary A and October, February, and May for all other levels. No other descriptive information is provided.

Reliability: Estimates computed by the alternate forms method as well as by the split-half procedure, range from .67 to .94. In the main, the estimates for the subtests are above .80, but some of them are so low that extreme caution must be used in interpreting subtest score differences. Information concerning the reliability samples is lacking.

Validity: Validity evidence as such is not presented. While the authors describe the population used to determine which items were to be retained, neither descriptive data concerning the tryout sample nor information concerning the sources studied to develop items are described. Correlations reported for each of grades 4 to 8 between the subtest scores and the Lorge-Thorndike Verbal IQ range from .18 to .86.

• *Comment:* The content appears to reflect current trends in the teaching of reading as well as in recognizing that the experiential domain of today's student is much broader than it used to be. Some of the subtests are so highly speeded that many students do not finish in the allotted time. Some of the reliabilities of the subtest scores are so low that extreme caution must be used in interpreting subtest score differences and in evaluating gain scores on subtests. Despite the criticisms noted here and above, the tests are popular with teachers, who appear to think they serve adequately the purposes for which they were designed; they may not do this.

San Diego Quick Assessment. Unpublished.

Grade Levels: 1-6+

Content: Graded list of 130 words taken from basal reading series and the Thorndike Word List. Words are presented on index cards in 11 sets of 10 words each.

Scores: No scores -- instructional level in reading is determined by level of last card correctly read.

Norming: No norms

Reliability: No information

Validity: No information

• *Comment:* This test was developed by two professors at the California State University at San Diego and is used in their classes in methods of teaching reading. It is not a standardized test. It is a highly subjective test since the examiner decides whether a word has been correctly read; and if not, why not. Its usefulness in any school situation would be limited to the ability of the examiner to make just the right decisions. It is definitely not recommended for class grouping.

DIFFERENTIAL APTITUDE TESTS

Differential Aptitude Tests (DAT). Psychological Corporation, Revised 1972.

Grade Levels: 8-12 (1 level, 2 forms)

Raw Scores: 9 subtest scores and a VR+NA score: verbal reasoning (VR), numerical ability (NA), abstract reasoning (AR), clerical speed and accuracy (CSA), mechanical reasoning (MR), space relations (SR), language usage I: spelling (SPEL), and language usage II: grammar (GRAM).

Derived Scores: Percentile ranks and stanines for each of the eight subtests and for the combined raw scores on VR and NA

Norming: The norms group included more than 64,000 students from 76 school districts in 33 states and the District of Columbia. Separate sex and grade level (8-12) norms are provided. The testing of the normative sample was done in the fall. However, the authors also provide spring norms for grades 8-11. These spring norms were obtained by interpolating between the fall norms of successive grades. The accuracy of these interpolated norms is debatable.

Reliability: Split-half reliability coefficients computed separately for each sex and each grade are reported for both forms for all subtests except CSA. The mean reliability coefficients for the separate subtests range from .79 to .97 for boys and from .80 to .97 for girls. MR was the least reliable subtest for girls and CSA the least reliable for boys. For each grade, standard errors of measurement were computed on each subtest for boys and girls -- these range from about 2.3 to 5.5 raw score points. The correlations between grade 9 and grade 12 scores, on a long-term basis, range from .60 to .86. MR and CSA are the least stable subtests. VR is the most stable.

Validity: The research on the prediction of course grades is summarized according to subject areas. Median correlations (across all studies) between the best subscores on the DAT and the criterion course grades range from the upper .40s to the low .60s. However,

all four major subject-matter areas can be predicted successfully using the same score: VR+NA; thus, the differential validity of the DAT in predicting course grades is not very well substantiated. The prediction of achievement test results follows essentially the same pattern as the prediction of course grades. Concurrent validity studies showing the correlation between the VR+NA score and tests of general intelligence range mostly in the .70s and .80s, as high as the correlations between most tests of general intelligence. It certainly appears that the VR+NA score serves the same purpose as general intelligence test scores.

• *Comment:* The primary suggested use of multifactor aptitude tests has been for educational and vocational guidance. The administration of the DAT in grade 8 or 9 can provide information that is relevant to the decisions a student must make concerning future educational plans. The subtests predict a variety of criteria, and the descriptive value of the subtest scores is not to be underemphasized. Many counselors are appreciative of the fact that students who would perform at a low level on a test of general intelligence may do well on some of the subtests of the DAT; thus, the counselor can say something of a positive nature concerning the student's abilities, and the student leaves the counseling interview with a better self-concept than if one could only intrepret the low scores on a general intelligence test. The DAT is intended for help in decision-making regarding high school curriculum on the part of student, parents, teachers, counselors combined. It is not intended that such decision-making be made at too early an age or that it necessarily be permanent.

APPENDIX D

A GLOSSARY OF MEASUREMENT TERMS USED IN THIS HANDBOOK

The following are brief definitions of the terms most frequently used in discussing tests and the results of testing. The definitions are intended to provide educators and community leaders with some understanding of the implications of the terms for the tests with which they are concerned without burdening them with technical details.

Ability test: also known as aptitude test or intelligence test. An ability test attempts to measure the combination of native and acquired abilities needed for school work, that is, the academic potential of the student. Theoretically, the items on an ability test are based on research into the human learning process while items on an achievement test are closely related to specific classroom teaching and learning. In practice, there is often little difference between the two tests.

Achievement test: a test that attempts to measure the extent to which a person has mastered certain specific skills taught in the classroom. Achievement tests are frequently administered in "batteries," or groups, of from four to 10 separate tests covering different aspects of the curriculum. All batteries have subtests in the specific areas of reading, arithmetic skills and problem solving, and language skills, and some batteries go beyond the above-mentioned fields and attempt to test achievement in areas such as science, social studies, and study skills. For the upper grade and secondary school levels, separate achievement tests are available in most every subject area. Achievement tests do not have to be commercially published. Applied loosely, the term "achievement" can be used to describe any test a teacher gives to find out how well students have learned a particular subject. (See CRITERION-REFERENCED TEST, DIAGNOSTIC TEST, INVENTORY TEST.)

Average: a measure of central tendency. The three most widely used averages are the median, the mean, and the mode. (See MEDIAN, MEAN, MODE.) When the word "average" is used alone, it generally refers to the mean. A score should be called "above average" or "below average" only after the type of average being used is identified. If the average used is the mean, for example, and the mean's value is 60 for a particular 100-point test, then the scores 61-100 are "above average" and the scores 0-59 are "below average." If, on the other hand, the average used is the median, the top 50 percent of the scores are "above average" and the bottom 50 percent are "below average."

Battery: a group of tests -- usually from four to 10 -- standardized on the same sample population. Because each test in the battery has the same set of norms, the tests are easily comparable with one another. For example, the reading skills of a student can be compared to his other skills. Almost all achievement tests used at the elementary school level and many of those designed for use at the high school level are given in batteries. (See NORMS for an explanation of how a test is standardized.)

Composite score: a total score that combines several subscores, usually by averaging.

Converted score: a general term referring to any of a variety of "transformed" scores, in terms of which raw scores on a test may be expressed for such reasons as making interpretation easier and permitting comparison with scores on other test forms. Percentile scores, scaled scores, and stanines are all converted scores.

Correlation: the "going-togetherness" of two scores or tests. If there is a tendency for students with high IQ scores also to be high in reading ability, then there is a high correlation between scores on IQ tests and scores on reading tests. This does not mean that having a "high IQ" is the cause of a person's high reading score; it merely means that there is a statistical relationship between the two phenomena.

Criterion: a requirement which a test must meet. Before a test can be considered valid, it must meet certain criteria. Suppose a test is designed to measure a student's understanding of the multiplication and division of fractions. If it is shown that the test does not accurately gauge a student's understanding of these two arithmetic processes, then it has not met its criteria; it is not valid. Thus, a criterion is a standard used to test a test. (See VALIDITY.)

Criterion-referenced test: an achievement test that ideally covers small units of content. Its content is closely related to what has been taught in the classroom. Since most criterion-referenced tests are developed according to local specifications, they are rarely standardized on a national norms group.

Culture-fair test: a test limited in content to that which is common to all cultures. Regardless of the culture in which a child has been brought up, a culture-fair test should measure "fairly" his capacity for "knowing."

Culture-free test: There is no such thing as a culture-free test. Such a test could measure only "inherited" abilities. While this may not seem impossible in theory, in practice a test cannot be made culturally sterile. The language structure used in the test and the background knowledge required to answer questions would both be culturally determined. Even so-called "performance tests," which do not use language, cannot eliminate cultural bias, though they may reduce it. Furthermore, even as an infant a person has absorbed to some extent the culture in which he is brought up and this will influence his responses to test items as well as to the conditions under which the test is taken.

Diagnostic test: an achievement test used to "diagnose" or analyze, that is, to locate an individual's specific areas of weakness or strength and, wherever possible, to suggest their cause.

Grade equivalent: a numerical rating which indicates what average level of achievement a given score represents. Grade equivalents are calculated on the performance of the norms group and based on a 10-month school year. Thus, if a child received a 5.7 grade equivalent, his test score would be equal to the average test score of the children in the norms group who were in the seventh month of grade 5.

Intelligence Quotient (IQ): originally, the ratio of a person's mental age to his chronological age: MA/CA, multiplied by 100 to eliminate the decimal. This quotient rating has fallen into disrepute and has gradually been replaced by the deviation IQ (DIQ) concept, which is based on the difference (or deviation) between a person's score and the average score for persons of his chronological age. Though the terms "superior," "above average," "average," and "below average" have been used to label IQ scores, it is perhaps useful and less misleading to interpret an IQ score by comparing it with national scores. Figure I is an illustration of such a comparison. It relates deviation IQ scores to percentile ranks.

Inventory test: an achievement test that atempts to cover thoroughly a small unit of instruction in specific subject matter in order to take "inventory" of an individual's stock of knowledge. (It is often called a pretest.)

Mean: the sum of a set of scores divided by the number of scores.

Median: the point at which half the scores in a group fall below and half above, unless the median itself is one of the scores. The median is always the 50th percentile. Some people who misunderstand the term are dismayed that half of the children who have taken a certain test have fallen below the median. They are unaware that this is always the case. Even if no one missed more than a few questions, half of those taking the test would still receive scores below the median.

Mental age: the age for which a given test score on a mental ability test is normal or average. For example, if the average score of children who are 6 years, 10 months, of age is 55 on a particular test, then a child making a 55 on that test is said to have a mental age of 6-10, regard-

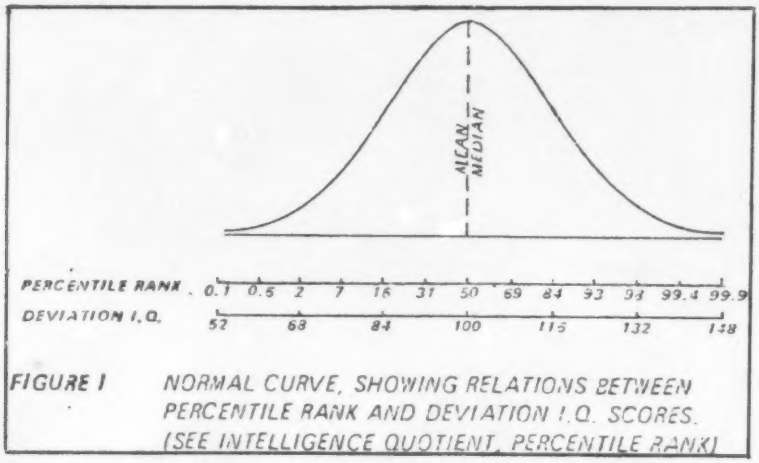


FIGURE I NORMAL CURVE, SHOWING RELATIONS BETWEEN PERCENTILE RANK AND DEVIATION I.Q. SCORES. (SEE INTELLIGENCE QUOTIENT, PERCENTILE RANK)

less of his chronological age; that is, he could actually be 8 years old and still have a mental age, according to the test score, of 6 years, 10 months.

Mode: the score or value that occurs most frequently in a distribution. In the distribution 1, 2, 3, 4, 4, 4, 5, 7, 9, the number 4 is the modal value.

Multiple-choice item: a test item in which the examinee's task is to choose the correct or best answer from several options. An example: Chicago is (a) a town, (b) a city, (c) a state, (d) a country.

Normal distribution: a distribution of scores or measures that in graphic form has a distinctive bell-shaped appearance. This curve is known as a "normal" curve. In such a distribution, scores are distributed symmetrically above and below the mean, and the mean and median are the same. (Figures I, II, and III are examples of normal distribution curves.)

Norming or Standardization: Suppose norms are to be developed for Test X, which will eventually be given to all school children in public elementary schools. The publisher will first try to select a norms population that is representative of all public school children, although it may comprise only 5 percent of that total population. Test X is given to this group and the results become norms. Test X is now said to be standardized on this norms population. When Test X is given to elementary school children, wherever they are located and no matter to what socioeconomic group they belong, it should be possible to compare the results obtained to these norms. National norms are simply norms that have been derived from testing a group selected from all over the United States. Most standardized tests use national norms. Some use regional norms or norms based on students from different socioeconomic groups as well.

Norms: statistics that give meaning to a test score. A raw score can have meaning only when it is referred to some type of group or groups. Suppose a man is 76 inches tall. Is he above or below normal height? For a Watusi, he would be about average; in Japan, he would be far above average. Similarly, in order to label a score above or below average, a set of norms must be established against which the score can be compared. In other words, a test must be standardized on a norms population before scores obtained on that test can be interpreted. Norms can be reported in several ways, including percentile ranks, stanines, mental age, grade equivalents, and IQ scores. (These terms are defined elsewhere in the Glossary.) These norms systems are different ways of expressing the same thing. For example, a 9-year-old child of "slightly above average intelligence" could rank in the 63rd percentile and the 6th stanine in relation to other 9-year-olds; have a mental age of 10, a grade equivalent of 5.1, and an average of 109. "Norms" is the term used to describe the full range of scores obtained by the norms population, while the "norm" refers only to the midpoint or average of that range. A norm is merely a statement of the average found at a particular time by test publishers.

Objective test: a test made up of items for which correct responses are set up in advance; scores are unaffected by the opinion or judgment of the scorer. For example: True or False: Washington, D.C., is

the capital of the United States. An objective test is in contrast to a subjective test such as an essay examination, to which different test correctors may assign different scores.

Parallel forms of tests: two or more forms of a test that are assembled as closely as possible to the same statistical and content specifications so that they will provide the same kind of measurement at different administrations.

Percentile: a point in a distribution. A score coinciding with the 35th percentile equals or surpasses the score of 35 percent of the persons in the norms group and equals or falls below 65 percent of the performances in the group. Percentile has nothing to do with the percent of correct answers.

Percentile band: a range of percentile ranks which takes into account the measurement error that is always involved in assessing raw test scores. A student might be told that his score fell within the 60th to 75th percentile band rather than at a particular percentile rank, say the 73rd percentile, because it is possible that on another day, taking the same test, his score may rank anywhere from 60 percent to 75 percent. The errors of measurement that are inherent in any test make the reporting of test scores in band form desirable.

Percentile rank: the percent of scores in a distribution equal to or lower than a particular obtained score. A percentile rank should not be confused with the percent of correct answers an examinee has obtained on a test.

Performance test: a test involving some manual response on the examinee's part, generally a manipulation of concrete equipment or materials. There are many types of performance tests, but they all have one characteristic in common: they do not depend upon verbal symbols alone to measure individual differences. The role of language is excluded or minimized.

Power test: a test intended to measure level of performance unaffected by speed of response; there is either no time limit or a very generous one. Items are usually arranged in order of increasing difficulty. This type of test is in contrast to a speeded test, which has a definite time limit and may penalize the slow-working student.

Random sample: a sample taken from some total population (for example, American children in elementary school) in such a way that every member of that population has an equal chance of being included. For example, drawing names out of a hat in which each person's name has been put only once gives a random sample. By such a "blind" process, testers hope to avoid choosing one "type" of person more than another "type" and to select a norms population much smaller than the total population it represents. In a stratified random sample, which has recently become more common, those chosen are representative of specified subgroups of the total population. For example, you might have a separate hat to draw out of for each grade, each ethnic group, each geographic area, and/or each socioeconomic status. This process is more "biased" than pure random sampling in that it restricts the groups it can choose from.

Range: the difference between the lowest and highest scores obtained on a test by some group.

Raw score: the first quantitative result obtained in scoring a test. This could be the number of right answers, the number of right answers minus some fraction of the number wrong, the number of errors, or some similar unconverted, uninterpreted measure.

Reliability: the extent to which a test is consistent in measuring whatever it does measure; the dependability of a test, its relative freedom from errors of measurement. If you weigh yourself on the same bathroom scale several times with the same result, then the scale is assumed to be reliable. Similarly, if the same test or a parallel form of the test yields approximately the same results from administration to administration, the test is assumed to be reliable.

Representative sample: a sample that corresponds to or matches the population of which it is a sample with respect to characteristics important for the purposes under investigation.

Scaled score: the score on a test when the raw score obtained has been converted to a number or position on a standard reference scale.

Speeded test: (See POWER TEST.)

Standard deviation: a measure of the spread of a score distribution. Calculating the standard deviation of a set of scores shows how they deviate from the mean score. In general, the smaller the standard deviation, the closer the scores will group around the mean; the larger the standard deviation, the more spread there will be. The norms statistics for most standardized tests are based on relatively large standard deviations of between 10 and 20. For a normal distribution (which is the most common distribution) there is an exact relationship between the standard deviation and the percentage of cases falling within each standard deviation from the mean. Figure II shows this relationship.

Standard error of measurement (S.E. Meas.): a statistic which gives the possible magnitude of "error" present in a score. S.E. Meas. indicates the amount a given score may differ from its hypothetical "true" score. The standard error is an amount such that about 2 times out of 3 the "obtained" score would not differ by more than one standard error from the "true" score.

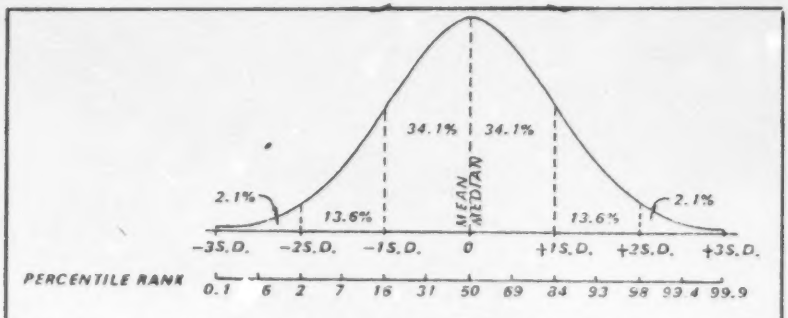


FIGURE II NORMAL CURVE, SHOWING THE PROPORTION OF CASES THAT ALWAYS FALL WITHIN INDICATED STANDARD DEVIATIONS FROM MEAN. (SEE STANDARD DEVIATION)

Standardized test: a test prepared by specialists, administered according to uniform directions, scored in conformance with definite rules, and interpreted in terms of certain normative information. Reliability and validity data are usually provided. Such tests are commercially published and for general use.

Stanine scale: a nine-point scale used to help interpret [illegible]; a norms system. The stanine (short [illegible] standard-nine) scale has values from 1 to 9, with a mean of 5. Each stanine corresponds to a certain range of percentiles (See Figure III). Each stanine (except the two extremes 1 and 9) has the same width, thus dividing a score distribution into seven equal parts within the two extremes. Stanines simplify reporting score results and are usually used in reporting national norms statistics.

True score: a score entirely free of error; hence, a hypothetical value that can never be obtained by testing, which always involves some measurement error. A true score is sometimes thought of in this way: if a student took a certain test an infinite number of times (assuming no practice effect or change in the student) the average score of this infinite number of scores would be the true score. (See STANDARD ERROR OF MEASUREMENT.)

Validity: the extent to which a test does the job for which it is used. A bathroom scale is a "valid" measure of weight but a ruler, for example, is not. Similarly, a reading test would not be a valid measure of achievement in writing.

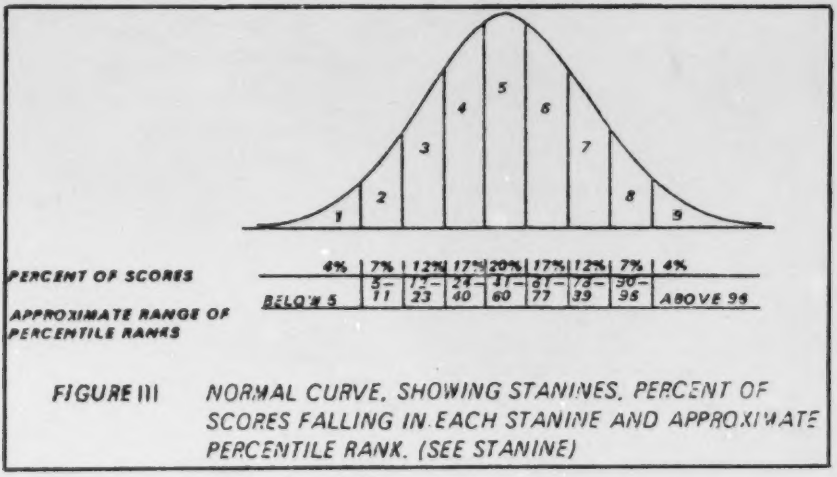

1
2
3
4
5
6
7
8
9
PERCENT OF SCORES
4% | 7% | 12% | 17% | 20% | 17% | 12% | 7% | 4%
APPROXIMATE RANGE OF PERCENTILE RANKS
BELOW 5 | 5–11 | 12–23 | 24–40 | 41–60 | 61–77 | 78–89 | 90–95 | ABOVE 95

FIGURE III NORMAL CURVE, SHOWING STANINES, PERCENT OF SCORES FALLING IN EACH STANINE AND APPROXIMATE PERCENTILE RANK. (SEE STANINE)